It does not need to be voiced to be counted

Non-verbal behaviour influences assessors' global marking when examining medical students using objective structured clinical examinations

Sami Saad A. Alnasser

The University of Leeds
School of Medicine

Submitted in accordance with the requirements for the degree of PhD (Doctor of Philosophy).
Submitted for examination in January 2016

It does not need to be voiced to be counted

Non-verbal behaviour influences assessors' global marking when examining medical students using objective structured clinical examinations

Submitted in accordance with the requirements for the degree of PhD (Doctor of Philosophy).
Submitted for examination in January 2016

Contents

List of tables

List of figures

Acknowledgments

I would like to thank my parents for their great love and support. Although I was away from them during the time I spent in the UK, their love was always with me.

I would like to thank my sponsor and college in my home town. They gave me the chance to fully concentrate on my study, and they provided me with all help and support I needed.

I cannot forget to thank my supervisors, Professor Richard Fuller and Professor Trudie Roberts , for their outstanding supervision, support and cooperation. They created a very conducive learning environment for me. They have their own special way to inspire others which I have found very charming. It was a real pleasure meeting them and learning from them.

Abstract

Introduction:

Whilst OSCEs are a well-recognised format for assessing clinical competence, an increasing body of research focuses on the factors that contribute to differences in assessors' judgement in performance assessment. Perspectives from psychosocial research have explored factors influencing these differences, but less attention has been paid to non-verbal behaviours of candidates, assessors and patients that could influence assessors' judgements during OSCEs. This PhD report investigates how non-verbal behaviour influences assessors' global marking when examining undergraduate medical students using OSCEs.

Methodology:

In reaching theoretical saturation, 18 OSCE assessors participated in 1:1 interviews (11 male; 7 female, all medically qualified and had undergone OSCE faculty training). Each participant scored 2 videos of students consulting with a simulated patient (these were carefully constructed to layer in multiple non-verbal behaviour types), and made judgements on each performance using a standard scoring format and written feedback. A retrospective think aloud methodology was used as a stimulus to explore factors in the students' performances. Interview transcripts were coded and a modified grounded theory approach used to develop a framework to interpret results.

Results:

Thematic analysis revealed a rich framework where the interaction of non-verbal behaviours of assessors, patients and candidates all contributed to global ratings. Assessors' identification and response to candidate behaviours was complex and individual. Subthemes included the importance of 'body language' and the impact of assessor fatigue, coupled with individualistic approaches to the use of (and reliance on) pre-determined stereotypes. All these themes are further influenced by organisational and environmental factors.

Discussion:

In the 'theatre of performance' of the OSCE, all the characters contribute to variance – and thus (unlike many other papers) this research does not just focus on one character or another, but all and the environment. The nonverbal behaviours of the three 'characters' in the OSCE (student, patient and assessor), and the environment in which it is situated, make significant contributions to global ratings and contribute to the multiple factors that influence inter-rater reliability. This is important in station and scoring format design, assessor selection and training and the ongoing research into assessor decision-making in high stakes performance tests.

Conclusion:

'It does not need to be voiced to be counted'. Non-verbal behaviours within an OSCE station have significant impact on assessor judgements, and contribute to the multiple factors that reduce inter-rater reliability.

Chapter 1 Introduction

Assessment is a part of almost all daily life activities. For instance, buying a shirt or pen requires assessment. The buyer will compare one item with many others in order to decide which one is more suitable or appropriate. Many factors will play a role in this process before making a final decision about the appropriate item. For example, price, quality, or colour are just some criteria that might be taken into consideration before buying such items. Likewise, assessors make decisions about learners' knowledge, skills and attitudes, applying certain related criteria. Therefore, and from an educational perspective, assessment is the process of defining, selecting, designing, collecting, analysing, interpreting, and using information to increase students' learning and development (Erwin, 1991). Assessment consists of the activities undertaken by assessors -and by their students in assessing themselves- which provide information to be used as feedback to modify teaching and learning activities (Black & Wiliam, 1998). The previous definitions emphasise the importance of using and utilising assessment in enhancing students' learning. This trend in enhancing the learning process of students through the usage of assessment is different from typical assessment functions where the assessment was mainly used to test existing knowledge.

The process of assessing learners can be conducted, for example, by different means or types of assessments such as paper and pencil test in the classroom, for declarative knowledge, or observing students in the clinic for skills and attitudes examination. The purpose of the assessment will decide which means is more appropriate as will be discussed later. Nevertheless, assessment of learners may not be accomplished appropriately and optimally due to several challenges related to validity, reliability and

other factors that can influence the output of any assessment process. These possible challenges associated with assessing learners can negatively affect the learning process and lead to unpleasant ramifications with regard to the quality of teaching, learning and graduation of learners. Therefore, assessment in education has been studied and discussed thoroughly during the last few decades in order to overcome current existing drawbacks of assessing learners. Assessment will always be an important element of any educational system because it has a role in driving learning and filling gaps in instruction and the curriculum (Miller, 1990). Assessment instruments work together with content, teaching, learning activities, and evaluation in order to develop optimal curricula (Prideaux, 2003).

The importance of assessment increases when it comes to examining and graduating medical students. Medical schools are responsible for graduating qualified medical doctors who will be able to take the responsibility of dealing with vulnerable patients and providing them with medical care to fulfil the obligations placed on medical schools by society. Since the 1950s, there has been rapid and noticeable change in the way assessment is conducted in medical education with more focus on assessing clinical skills such as physical examination, communication skills, procedural skills and professionalism (Norcini & McKinley, 2007). However, and regardless of the major developments in the assessment process taking place in medical education, ‘‘there is probably more bad practice and ignorance of significant issues in the area of assessment than in any other aspect of higher education’’ (Boud, 1995, p. 35). Long cases, for example, are unlikely to produce an accurate reflection of ability, and were criticised with regard to their reliability and validity as the learner was not observed communicating with the patient (Harden et al., 2015). The clinical examination was described by Stokes (1974) as the ‘‘half-hour disaster session’’ and the ‘‘sacred cow of British medicine’’

(Harden et al., 2015). Therefore, assessment in medical education has been an active field for investigation and development as meaningful assessment is a prerequisite for enhancing the medical care provided to society.

This research began by recognising the need for understanding some issues related to reliability to help enhance assessment in medical education. The context was chosen to be the objective structured clinical examination (OSCE) due to the popularity of this assessment instrument in many countries around the world (see Appendix 1 for how the OSCE works in Leeds). Although OSCEs are one of the most common performance assessment tools, they can be subject to a variety of potential threats to their reliability. The OSCE, as will be described later in detail, utilises human observation to inform assessment. However, taking advantage of human observation to inform assessment of its assessors and learners has faced challenges in medical education (Gingerich et al., 2011). Research identifies this challenge in that rater-based assessments generally reveal psychometric weaknesses (Albanese, 1999; Kassebaum & Eaglen, 1999; Lurie et al., 2009b; Williams et al., 2003) including measurement errors of leniency (Cacamese et al., 2007), undifferentiation (Silber et al., 2004), range restriction (Hatala & Norman, 1999), bias (van Barneveld, 2005), and unreliability (Clauser et al., 1999). Unfortunately, such psychometric weaknesses have not been adequately solved (Yeates et al., 2015) through either reformulation (Cook & Beckman, 2009; Donato et al., 2008) or training (Cook et al., 2009; Crossingham et al., 2012; Holmboe et al., 2004). Consequently, 'assessor cognition' has been researched to comprehend cognitive aspects causing these limitations.

This research aims to understand some issues related to inter-rater reliability by investigating how non-verbal behaviour influences judgements. In order to understand the decision making process, the underlying contextual factors need to be studied and

understood (Hoffman et al., 2004). The assessor's decision processes in the OSCE setting have not been adequately studied. Rather, some comparative research has been conducted (Benner, 1984; Benner et al., 1996; Benner & Tanner, 1987; Cooper & Bond, 2006). Therefore, researchers have found a different way of investigation and research, that is about studying how assessors make judgements and what could affect their assessments from different angles; social, cognitive, psychological and medical. ''Clinical medical examinations are subject to a variety of potential threats to their reliability. While candidates' scores vary according to their ability, leading to true variance in their scores, error variance can result from a variety of sources'' (Denney et al., 2013, p. 718).

Whilst differences in assessor judgements have been labelled as 'error variance', recent workplace assessment based research has explored different perspectives of assessors (and their decisions) through constructivist lenses, identifying them as 'trainable, fallible or meaningfully idiosyncratic' (Gingerich et al., 2014). Perspectives from psychosocial research have explored factors influencing this idiosyncrasy, but less attention has been paid to non-verbal behaviours of candidates, assessors and patients that could influence assessors' judgements during OSCEs. Therefore, this research investigated how non-verbal behaviour influences assessors' global marking when examining undergraduate medical students using objective structured clinical examinations.

Chapter 2-A Meaningful assessment and the OSCE

Introduction

The assessment of learners in medical education is one of the recurrent matters that leads to debate about the extent to which assessment boosts or undermines learning (Hays, 2008). Foster and Cone (1995) noted that ‘‘science rests on the adequacy of its measurements. Poor measures provide a weak foundation for research and clinical endeavors’’. Credible and meaningful assessment is essential to help reduce the number of incompetent healthcare practitioners (Shanley, 2001).This chapter looks at what makes assessment meaningful with greater focus on the context of this research, the OSCE.

Before I decided what assessment method I should select and use, it was essential to understand what I wanted to assess. ‘‘To assess students’ competence what we need is to observe their skill. Whilst this may seem obvious, all too often in medicine we fall into the trap and rely on testing the students’ knowledge with written assessments when what we are interested in is their clinical competence. This represents the bottom of Miller’s Pyramid, shown in Figure 1 (Miller, 1990), at the ‘knows’ and ‘knows how’ levels rather than the ‘shows how’ level’’ (Harden et al., 2015, p. 6). Competence should not be seen as an achievement, rather it is a habit of lifelong learning (Leach, 2002). It is contextual and reflects the relationship between a candidate’s abilities and what he or she is required to perform in a particular situation in the real world (Klass, 2000). Professional competence includes the accustomed and careful usage of communication, knowledge, technical skills, clinical reasoning, judgement, emotions, values and reflection in daily practice (Epstein & Hundert, 2002).

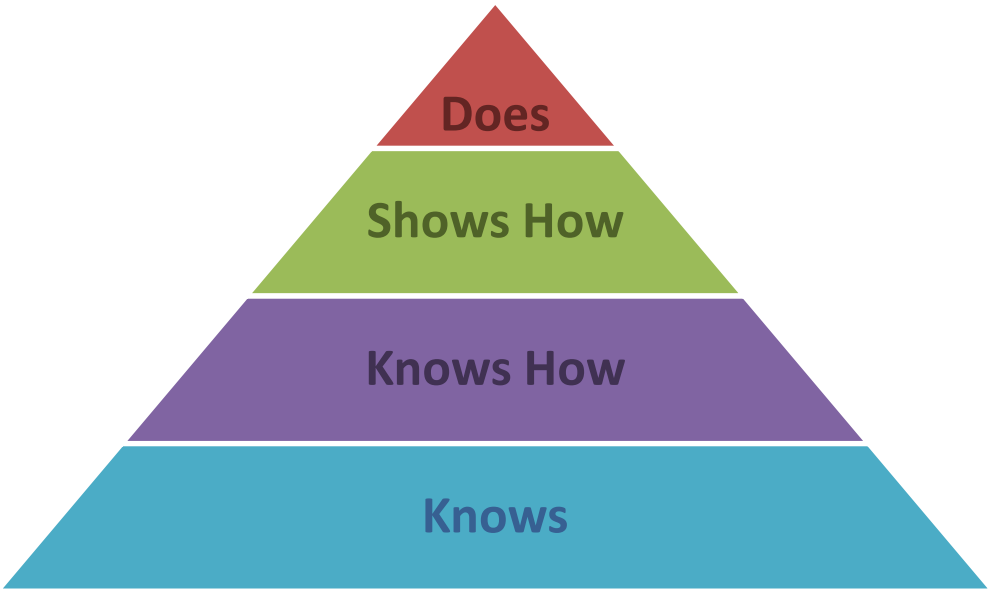

Figure1 Miller's Pyramid (Miller, 1990)

Medical knowledge and clinical skills used to be habitually assessed utilising written or oral examinations (Norcini, 2005). However, health care has become increasingly complex which necessitates and requires conceptualisations of competence as collective, situated and dynamically produced through social interaction (Lingard, 2012). Competence, or what the student is able to do, is ideally assessed to provide insight into actual performance, or what is habitually done when not observed, as well as the capability to adapt to change, locate and generate new knowledge, and develop overall performance (Epstein, 2007; Fraser & Greenhalgh, 2001).

From interpretivist and social-constructivist approaches' point of view (Delandshere, 2002; Johnston, 2004; Koch & DeLuca, 2012; Moss, 1996; Moss et al., 2006), performance assessments have been perceived as social constructions or interpretations, rather than absolute, objective truths (Johnston, 2004). In other words, there is no single 'accurate' score or 'objective' rating of performance (Govaerts & van der Vleuten, 2013).

New efforts have been made to provide accurate and reliable assessments of the competence of candidates in medical schools and training programs over the past decades (Batalden et al., 2002; Epstein & Hundert, 2002; Leung, 2002). Every assessment tool has its own strength and weakness. The utilisation of multiple observations and assessment tools is a strategy that could help confront such inevitable flaws when assessing candidates (Epstein & Hundert, 2002; Wass & van der Vleuten, 2004). In addition, the combination of knowledge, skills and behaviours cannot be assessed by a single assessment method. Therefore, Epstein (2007) argues for the utilisation of a blend of assessment methods to assess a variety of learning domains. This combination of assessment methods can be called a 'test battery' approach (Hamdy et al., 2010), and the OSCE is considered an essential examination in this test battery in the assessment of clinical performance in a simulated experience (Khan et al., 2013). Furthermore, different types of assessment methods can be incorporated within the OSCE format (Hodder et al., 1989; van der Vleuten & Swanson, 1990) such as written or oral questions. ''The examinee's response may be in the form of a multiple choice question (MCQ), a short constructed response to a question, a note about the patient they have seen – sometimes called a 'post-encounter note', a letter referring the patient for further investigation or treatment, or an oral report to an examiner'' (Harden et al., 2015, p. 7).

The OSCE was first described by Harden and Gleeson in 1975 as a new assessment tool that could replace the old existing assessments of clinical performance (Harden et al., 1975). It was designed to solve some issues with the validity and reliability, which will be described later in more detail, of clinical performance assessment methods such as the long and short case assessments. In addition, the OSCE was introduced to overcome other problems with the traditional clinical examination such as the small sample of skills

examined and the subjectivity or bias associated with the examiner's rating of the candidate (Harden et al., 2015). Since then the usage of the OSCE has become popular, see Figure 2, as an assessment method within both undergraduate and postgraduate clinical education (ibid).

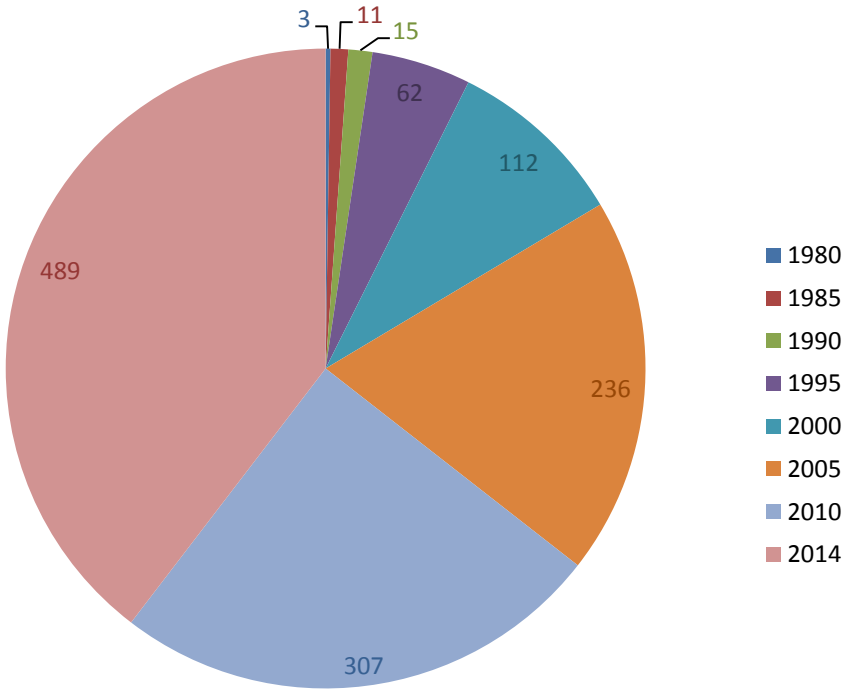

Figure 2 The increasing popularity of the OSCE shown by the increasing number of papers published over the last four decades (the number of papers listed in PubMed, http://www.ncbi.nlm.nih.gov/PubMed) from (Harden et al., 2015).

''The OSCE is a performance-based examination in which examinees are observed and scored as they rotate around a series of stations according to a set plan. Each station focuses on an element of clinical competence, and the learner's performance with a real patient, a simulated patient, a manikin or patient investigations is assessed by an examiner'' (Harden et al., 2015, p. 1). The OSCE was introduced and designed as a novel assessment method, which could help the assessment of learners' clinical skills, attitudes, problem-solving and application of knowledge in one examination (Harden et al., 1975).

This assessment method is based on the principles of objectivity and standardisation, see Appendix 2, in which the students move through a series of time-limited stations in a simulated environment (Khan et al., 2013). The principles of objectivity and standardisation in the OSCE helps enhance the assessment of learner's performance against standardised scoring schemes by trained examiners (ibid). It is worth noting that each station has a different examiner, so students are assessed by a large number of people, in contrast to other forms of examinations (Harden et al., 2015). Physicians or other knowledgeable health professionals are the most common OSCEs assessors because they are required to determine whether the correct information was used by the student and the listed diagnoses were probable (Tamblyn & Barrows, 1999).

Table 1 Examples of clinical skills assessed in an OSCE (Harden, 1988)

Skill	Action	Example
History taking	History taking from a patient who presents a problem	Abdominal pain
	History taking to elucidate a diagnosis	Hypothyroidism
Patient education	Provision of patient advice	Discharge from hospital following a myocardial infarction
	Educating a patient about management	Use of an inhaler for asthma
	Provision of patient advice about tests and procedures	Endoscopy
Communication	Communication with other members of healthcare teams	Brief to nurse with regard to a terminally ill patient
	Communication with relatives	Informing a wife that her husband has bronchial carcinoma
	Writing a letter	Referral or discharge letter
Physical examination	Physical examination of a system or part of the body	Hands of a patient with rheumatoid arthritis
	Physical examination to follow up a problem	Congestive cardiac failure
	Physical examination to help confirm or refute a diagnosis	Thyrotoxicosis
Diagnostic procedure	Diagnostic procedure	Ophthalmoscopy
Interpretation	Interpretation of findings	Charts, laboratory reports or findings documented in patient's records
Patient management	Patient management	Writing a prescription
Critical appraisal	Critical appraisal	Review of a published article or pharmaceutical advertisement
Problem solving	Problem solving	Approach adopted in a case where a patient complains that her weight as recorded in the hospital was not her correct weight.

The features that characterise an OSCE can be highlighted as the ‘eight Ps’. (Harden, 1992; Harden et al., 2015, p. 9).

1- **Performance assessment**: the OSCE may be identified with a move from theory to practice. Examinees are assessed not just on what they know but also on what they can do.
2- **Process and product**: here the learner’s technical skills are assessed, for example, how they take a history, how a patient is examined or how the learner carries out a practical procedure. The learner’s findings, the results and their interpretation can also be assessed.
3- **Profile of learner**: The OSCE not only provides a single global rating for the learner but can also present a picture of his/her strengths or weaknesses in the different learning outcome domains.
4- **Progress of learner**: The OSCE assesses the learner’s progress during the curriculum and training programme and provides feedback to the learner and teacher as to strengths and weaknesses in the learner’s performance.
5- **Public assessment**: In the OSCE there is transparency as to what is being assessed. A discussion about what is assessed in an OSCE can lead to clarification of aims and expected outcomes relating to the course.
6- **Participation of staff**: Examinees are seen by a number of examiners, and staff from different specialties and healthcare professions can participate as examiners in the OSCE.
7- **Pressure for change**: The introduction of an OSCE can help to focus the learner’s attention on the competencies to be assessed. Poor overall performance in an

OSCE by a class of students highlights a need for a change in the education programme or a revision of the assessment.

8- **Pre-set standards of competence**: What is expected of a learner and the standard of performance appropriate for a pass in an examination are specified in advance.

Simulation

To ensure that candidates can demonstrate integration of prerequisite knowledge, skills, and behaviour in a realistic setting, the usage of simulations has been increasingly employed in medical education (Tekian, 1999). Both standardised patient simulations (e.g., Reznick et al., 1996; Whelan, 1999) and computer-based simulations (Clyman et al., 1999; Kneebone, 2003) have been commonly used to assess candidates' clinical skills and medical problem-solving in medical education, licensure and certification (Norcini & McKinley, 2007) to provide examinees with a simulated experience (Harden et al., 2015). Although simulated patient examinations have been successfully implemented in medical education, certification, and licensure, such examinations cannot always be used to assess all aspects of competence (Norcini & McKinley, 2007). For instance, the ability to perform procedures and manage life-threatening clinical situations need to be assessed. For such aspects of competence, the usage of computer-based simulations is more appropriate (ibid). In addition, the detection of abnormal physical signs is not always possible in simulated patients.

A simulated patient can be defined as a person competent to accurately and consistently represent a patient with a specific medical condition and is regularly incorporated into the OSCEs (Epstein, 2007). Based on an encounter between the standardised patient and a student, the quality of the candidate's performance (e.g., history-taking, interpersonal, and

communication skills) can be assessed by the simulated patient and assessor (Norcini & McKinley, 2007). For every patient scenario, a specific checklist is designed and generally focused on the candidate's ability to gather history and relevant information from the patient and perform the essential physical examinations (Tamblyn & Barrows, 1999). Simulated patient examinations have been widely used in medical schools and training programs to assess the ability of both physicians and medical students to gather medical history and physical examination data and establish a therapeutic relationship with the patient (ibid). The assessment of interpersonal skills could be conducted either by assessing candidates' interpersonal skills in every patient encounter, or by designing patient encounters that focus on the assessment of interpersonal skills (Norcini & McKinley, 2007).

Communication skills

Since this research is trying to understand how non-verbal behaviour influences assessors' judgements, it makes sense to briefly talk about communication skills in general, and non-verbal communication skills in particular. The need for good communication skills has been extensively discussed for its importance in medicine and other health professions (World Health Organization, 1993). Effective doctor-patient communication helps in improving patient satisfaction (Williams et al., 1998) adherence to therapy regimens (DiMatteo, 2004) and patient health outcomes (Stewart, 1995). All physicians require adequate communication skills to develop effective physician-patient relationships (Hall et al., 2004). Diagnostics, treatment, and prevention in primary health care always take place within the context of communication. Professional communication can be used to create confidence in the health worker and increase the likelihood of avoiding an erroneous diagnosis (Holte, 1990). Training of communication skills has

clearly enhanced the ability of learners in different medical and health fields to communicate with patients and establish rapport. For instance, training of communication skills has noticeably enhanced the ability of medical students (Quirk & Babineau, 1982; Rutter & Maguire, 1976; Schreier & Dub, 1981) and students of dentistry (Dunning & Lange, 1987) to communicate appropriately with patients in a way that could help gather data and establish rapport.

Communication skills have been defined in medical education as: ‘‘the interaction between doctors and patients (that) involves the forming of a relationship and the gathering and giving of information… to promote the physical, social and emotional well-being of patients and their families’’ (Adibi, 2014, p. 223). Furthermore, communication skills could refer to any communication between health professionals and patients or their relatives, or between health professionals and other colleagues. Communication skills can be written or oral, face-to-face, via telephone, electronic or via video transmission (Laidlaw & Hart, 2011). As a result, communication skills can either be verbal or non-verbal.

Non-verbal communication skills

Non-verbal communication skills should not be underestimated when it comes to assessing communication skills in general. Different studies have documented the central role of non-verbal communication skills in the medical encounter (Caris-Verhallen et al., 1999; Gorawara-Bhat et al., 2007; Gordon et al., 2006; Hall et al., 1995; Ishikawa et al., 2006; Larsen & Smith, 1981). Non-verbal skills are central to the development of rapport and trust between patients and health care professionals (Hall et al., 1995; Roter et al., 2006). In addition, doctors and patients are allowed, by non-verbal communication, to

gauge responses, to contextualise the meaning of verbal utterances, and to communicate a "hidden agenda" (Hall et al., 1981; Ishikawa et al., 2006). Intimacy, interest and balance of power have been shown to be conveyed by non-verbal communication (DiMatteo et al., 1980; Griffith et al., 2003; Larsen & Smith, 1981). Non-verbal communications are many and diverse. However, it is important to note that written letters or e-mails, for example, are considered non-verbal, but the focus here is on non-verbal communications that can be observed during an OSCE. Non-verbal communications have been classified differently by different researchers due to varying research purposes. To clearly make a distinction between verbal and non-verbal communication skills, the following table shows some examples of non-verbal behaviours that can be observed during an OSCE:

Table 2 Examples of non-verbal communication skills (Collins et al., 2011; Hall et al., 1995; Ishikawa et al., 2006)

Facial movements and expression	Gaze	Head movements	Body movements
Posture interpersonal distance	Angle of orientation toward other	Interpersonal touch	Voice
Nodding to facilitate patient's talk	Un purposive movements	Self touching	Speech rate
Match of tone and intonation with the verbal contents	Body position	Hand gestures	Affirmative gestures

Communication skills are no longer validly examined exclusively by traditional assessment tools such as written explanations (Vu & Barrows, 1994). Incorporating the assessment of communication skills into an OSCE has been recommended (Hodges et al., 1996) as a valid and reliable method (Colliver & Swartz, 1997; Epstein, 2007; Rushforth, 2007; Sloan et al., 1995; Sloan et al., 1996; Zubin et al., 2003). Global impression scales

or specific skills check lists are usually used to assess a learner's performance (Epstein, 2007). The following table shows an example of a global communication grading rubric.

Table 3 Global communication skills grading rubric (Schwartzman et al., 2011, p. 474)

GC skill category	1 Verbal expression-mechanics (HOW)	2 Verbal expression-content (WHAT)	3 Non-verbal expression	4 Interaction with patient/ health care professional	5 Organization & logic	6 Professional appearance & rapport
GC skill criteria	Speaks with proper grammar and fluency Uses correct pronunciation of Words Does not use filler words (e.g. Um, You Know, Like, Yeah, So) Speaks with appropriate rate of speech Uses appropriate volume of voice for the context Speaks with appropriate mod-ulation to effectively convey the message	Selects and uses vocabulary appropriate for the context Uses vocabulary appropriate for the audience Uses lay language when speaking to patients (avoids medical terms & abbreviations)	Maintains appropriate eye contact throughout the session with brief breaks to check products or notes when necessary Sits or stands in an upright position, demon-strating professional posture Does not engage in distracting gestures Does not create any awkward silences Maintains a comfortable physical distance based on the context	Displays active listening Displays empathy and sensitivity appropriate for the context Conducts the interaction in a non-formulaic manner Demonstrates perceptiveness by responding to cues and situations appropriately Shows respect and avoids speaking in hostile and condescending manner	Presents information in a logical order Information pre-sented flows smoothly with good transitions Shows flexibility/ ability to reorganize upon presentation or uncovering of unexpected information Maintains control of the session; shows ability to bring the conversation back to the topic when the audience detracts	Provides introductory greeting appropriate for the context (e.g. provides name and position only for new encounters) Attire and overall appearance is professional Ends the session in an appro-priate manner with proper closure
Score Scoring criteria	Scoring category	Category description				
	3 (Excellent)	Exhibits command by consistently meeting all criteria with minimal (<10%) deficiencies				
	2 (Satisfactory)	Satisfies many criteria but lacks consistency and some areas need improvement				
	1 (Needs improvement)	Does not satisfy several of the criteria or shows inconsistent delivery in many of the criteria				
	0 (Failure)	Failed to meet most of the criteria				

GC – global communication

Assessment strategies

Successful educational systems must create professional assessment strategies (Crossley et al., 2002). When conducted properly, assessments serve multiple purposes. These purposes (Amin & Khoo, 2003; Newble, 1998) include:

1- Determining whether the intended learning outcomes are met.
2- Supporting students learning.
3- Developing and evaluating teaching programs.
4- Understanding the learning process.
5- Predicting future performance.
6- Certification and judgement of competency.

Assessment requires careful consideration of many factors that can optimise and fulfill the goals of assessment and make it meaningful. The next section of this chapter describes how the OSCE is proven to be a meaningful assessment method.

Constructive Alignment

It is crucial for any education process to identify the current status of the student. This refers to the identification of the student's current knowledge and skills. Once this current status is identified, it is easier for educators and curriculum designers to set intended learning outcomes for a certain course or program. These intended learning outcomes are desired goals of any educational intervention. Teaching then would take place to achieve those intended goals. Different teaching and learning activities are applied and facilitated for the purpose of achieving preset aims and targets. However, what and how learners learn may depend to some extent on how they think they will be assessed (Biggs & Tang, 2007; Frederiksen, 1984; Newble & Jaeger, 1983).

Backwash is a term coined by Lewis Elton to refer to the effects assessment has on students' learning, to the extent that assessment might determine what and how students learn more than the curriculum intends (Elton, 1987). Therefore, backwash has the potential to either work positively or negatively on what and how students learn. For instance, if educators set assessment that mainly seeks rote learning, students will ultimately follow a surface learning approach. On the other hand, a deep learning approach is expected from students when assessment urges and promotes such an approach. As a result, backwash can work positively when the assessment is aligned to what students should be learning, intended learning outcomes, and teaching and learning activities (Biggs & Tang, 2007). Congruence in teaching and assessment (Figure.3) could help learners achieve such desired types of learning (Hays, 2008).

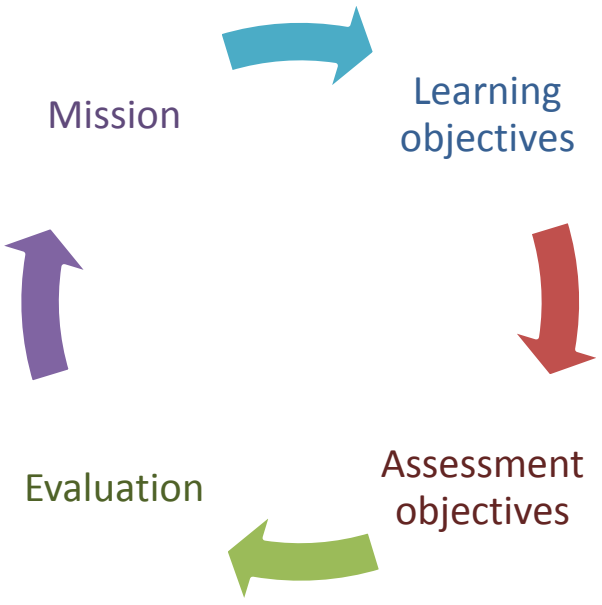

Figure 3 Congruence in teaching and assessment

The use of a blueprint has been commonly applied in the OSCE to help achieve constructive alignment. ‘‘A blueprint or a grid for the OSCE is prepared in advance. This outlines the learning outcomes and core tasks to be assessed at each station in the OSCE, for example, in the domains of communication skills, physical examination, practical procedures and analysis and reflection’’ (Harden et al., 2015, P. 5).

Table 4 Section of a blueprint showing content of an OSCE as tested at stations1,2,3,4,6,8,10 &12 (Harden et al., 2015, p.6)

	Body system				
Learning outcome	CVS	RS	NS	AS	ENDO
History taking	(2) Chest pain			(10) Diarrhoea	
Patient education					(1) Diabetes
Physical examination		(4) Asthma	(6) Hemiplegia		
Practical procedures	(8) BP	(12) FEV			
Problem solving			(3) Headache		
.					
.					
.					

AS, Alimentary system; BP, Blood pressure; CVS, Cardiovascular system; ENDO, Endocrine system; FEV, Forced expiratory volume; NS, Nervous system; RS, Respiratory system.

In addition, ‘‘a common and important comment from the students following an OSCE is that the examination is perceived as *fair*. One reason for this is that students in general see that the OSCE reflects the teaching and learning programme and the stations overall address the learning outcomes of the course’’ (Harden et al., 2015, p. 6). In OSCEs, stations are designed according to the expected learning outcomes of the specific stage of the curriculum (ibid).

Norm vs. Criterion Referenced Assessment

It is important before assessing students to have a preset standard that would function as a guide in differentiating between those who perform well and those who do not. This standard can be referred to as the systematic way of gathering value judgements, reaching consensus, and expressing that consensus on an examination either numerically or verbally. The credibility of such a standard, as it involves judgement, would vary depending on who sets the standards, the characteristics of the methods used, and the outcome (Norcini, 2003).

Educability, or the degree to which someone is educable, was considered to be a key in classifying who was bright and who was not, and therefore, education was seen as a device for sorting students and graduates out (Biggs & Tang, 2007). This perspective about education has caused a specific standard of assessment to be continually implemented where students can be sorted out and compared. This standard type of assessment is called norm-referenced assessment (NRA). The achievement differences between and among students are highlighted to produce a dependable rank order of students across a continuum of achievement from high achievers to low achievers (Stiggins, 1994). Consequently, the scores of tests are distributed normally and the grades of a given student are compared with the grades of other students (Norcini, 2003). In the OSCE, examinees are observed and scored as they rotate around a series of stations according to a set plan. This rotation of candidates might cause assessors to compare between them. However, the OSCE is designed to be more objective by ensuring that all candidates get the same exam and are compared against a certain predefined criteria.

Norm-referenced assessment might be useful for selecting a specific number of candidates, but it is definitely unacceptable for graduating medical students and clinical competency licensing (Wass et al., 2001). The reason behind this unacceptability is that under-achiever students may pass an exam if they get the highest grades among other students. In addition, students with acceptable grades may fail when other students are scoring very high marks. Hill and Parry (1994) comment that it is not difficult to place candidates in rank order, without being able to clarify what they are being put in rank order of. This need for greater clarity about the connection between the assessment and what it represents led, in the early 1960s, to the development of criterion-referenced assessments (William, 2000).

Criterion-referenced assessment, in contrast, has a different perspective in regard to the way of assessing students. While norm-referenced assessment ascertains the rank of students, criterion-referenced assessment determines what students can do and what they know, not how they compare to others (Anastasi, 1988; Green, 2002). Therefore, it looks at how well students are doing relative to a pre-defined performance level on a specified set of educational outcomes. The domain to which inferences are to be made is identified with great precision in criterion-referenced assessment (Popham, 1980). Likewise, the OSCE is based on the principles of objectivity and standardisation which help enhance the assessment of learners' performance against predefined standards and criteria and using standardised scoring schemes by trained examiners.The point here is to identify performances that tell assessors what has been learned, and how well. In contrast to the norm-referenced assessment, the failure rate in this second model may vary due to changes in students' abilities (Friedman Ben-David, 2000b). This kind of referencing is

the most relevant one for assessing students at university (Taylor, 1994) as one student's result is quite independent of any other student's (Biggs & Tang, 2007).

Contextualised Assessment

As mentioned earlier, assessment needs to be aligned with the intended learning outcomes. One of the advantages of this alignment is to decide whether the assessment needs to be contextualised or not. Assessment that is mainly looking for declarative knowledge can lead the students to perform in the abstract, out of context (Biggs & Tang, 2007). Examples of such an assessment include written exam or term paper. However, assessing only the lead-in declarative knowledge, not the functioning knowledge that emerges from it, is a common mistake (ibid). Treating learning as a product located in the mind of the student without paying much attention to the context where learning takes place has long dominated developments in instruction and assessment in medical education (Govaerts & van der Vleuten, 2013). It was believed that the nature of what is learned, or is to be learned, is somewhat independent of context (Hager, 2011). Therefore, and although many institutions believe so, competence should not be conceptualised as a stable trait, that once developed and established is considered to be portable and transferable from one context to another (Govaerts & van der Vleuten, 2013). Rather, there is an increasing body of research that challenges these conceptualisations of competence and professional performance. The role of social, cultural and organisational factors in shaping learning and performance development cannot be underestimated (ibid). Learning and expertise development are considered to be inseparably linked to features of the context in which the learning occurs, according to socio-cultural learning theories (Hager, 2011; Mann, 2011). Within-individual variation in performance is

significant and can be as large as between-individual variations (Deadrick et al., 1997; Fisher & Noble, 2004; Stewart & Nandkeolyar, 2007).

Therefore, and apart from assessing declarative knowledge, assessment requires contextualisation. Problem solving and cognitive skills are not generic (Epstein & Hundert, 2002; Norman, 2003) and performance in a specific problem area does not necessarily tell much about the performance of the student in other problem areas. For instance, performance of a student examining a diabetic patient may not have a strong correlation with the same student's performance dealing with a patient complaining of a middle ear problem. Assessors are required to realise that competence is contextual, reflecting the relationship between a candidate's abilities and the task required to be performed in a particular situation in the real world (Klass, 2000). Context specificity has urged that assessment of competence needs to consider more than one context. The competency of a student cannot be judged with confidence based on only one clinical encounter. Therefore, assessors have to employ many sampling strategies. This refers to the inclusion of multiple cases, multiple raters and multiple items in order to capture a wider image of the student's performance (Norman, 2003).

The OSCE, as mentioned earlier, was introduced and designed as a novel assessment method, which could help the assessment of learners' clinical skills, attitudes, problem-solving and application of knowledge in one examination (Harden et al., 1975) where the students move through a series of time-limited stations in a simulated environment (Khan et al., 2013). Since true intra-learner variation in performance could result from changes in the learner (e.g. due to fatigue, changing levels of competence or motivation) as well as changes in the assessment environment (Sturman et al., 2005), this fluctuation in performance might happen to medical students during OSCEs. Research findings in

medical education shows that context largely influences behaviours. For example, Durning and colleagues (2013) reported that contextual factors could influence clinical reasoning performance in ways that were related to the situation (OSCE) and participants (simulated or real patients) in the encounter (OSCE station) and the setting (Govaerts & van der Vleuten, 2013). It is important, however, to note that some essential skills, such as the ability to form therapeutic relationships, might be less dependent on content (Epstein et al., 2004).

The context in which performance is being assessed is a main difference between Work Place Based Assessment (WPBA) and the OSCE (Khan et al., 2013), not that the latter examines competence and the former performance, as is usually perceived (e.g., Boursicot et al., 2011). Since the performance of candidates on similar tasks can vary noticeably from one context to another, the difference between WPBA and OSCE is very significant (ten Cate et al., 2010). It is important to take into consideration that performance in the simulated environment might not transfer to actual practice settings (Norcini & McKinley, 2007). As a result, the OSCE needs to be seen as a method for assessing performance within a simulated environment (Khan et al., 2013). The performance of the learners in the OSCE may not be similar to their performance in the workplace on identical tasks (ibid). Furthermore, and unlike assessment in real life, it is not very applicable to assess non-clinical skills such as resource management, situational awareness, team working and leadership using the OSCE format because it mostly focuses on cognitive, psychomotor and affective skills, described as learning domains (ibid).

Analytic and holistic assessment

Analytic marking is one way of assessing a task by analysing it without looking to the whole picture. The task, an essay for example, is reduced to independent factors such as content, style, argument, format, and referencing. Each of which is rated on a separate scale and the final performance is assessed as the subtotal of the separate ratings (Biggs & Tang, 2007). In medicine, for example, a student carrying out a specific operation would be assessed, according to the analytic assessment approach, on his or her knowledge of anatomy, anaesthesia, asepsis and the performance skills essential for making clean incision. The aggregated mark may reach the minimum requirement for passing the assessment, but the student might remove the wrong part (ibid). Therefore, analytic assessment could produce unqualified surgeons by ignoring the overall performance and just focusing on the aggregation of marks achieved by the student. This strategy of analytic assessment is not the way things work in real life (Moss, 1994). However, analytic assessment can be noticeably used by assessors in giving detailed feedback about student's performance (Lejk & Wyvill, 2001).

OSCEs are usually marked in a tick box checklist rating format for each component of the examination. Nevertheless, checklists may neglect the general performance and holistic components of clinical competence (Cox, 1990). Therefore, global ratings of performance have been suggested as an advantage (Regehr et al., 1998). This global marking would give a general insight about whether or not candidates performed well in different domains and types of competence.

Convergent and divergent assessment

Students are not equal in their skills and creativity. This fact has raised the need to understand the nature of convergent and divergent assessment, and how both assessors and students can benefit from such assessment models. Treating every student the same when they are so obviously different from each other is the very opposite of fairness (Elton, 2005). This certainly does not mean that assessors have to differentiate between students in the way of teaching and assessing, but it urges the need to give a chance to more skillful and knowledgeable students to show their skills and knowledge. As discussed earlier in this chapter, intended learning outcomes are set before commencing the process of teaching and assessment. However, could assessment encourage unintended but desirable learning outcomes to emerge? The answer to this question is yes (Biggs & Tang, 2007) as will be explained. The terms convergent and divergent were originally coined by Guilford to describe two different forms of ability (Guilford, 1967):

Convergent ability, as in solving problems that have a specific and unique answer as in most ability test items. Convergent thinking is closed. Knowing a lot and getting it right should be only part of the academic story.

Divergent ability, as in generating alternatives, where the idea of being correct provides a way to other assessments of value, such as aesthetic appeal, usefulness, creativity and so on. Consequently, learning and competence should be considered expansive.

In medical education, some efforts to improve assessment appear to aim for the design of education and training that directs learner's learning in predictable ways, (Delandshere, 2002), as well as determining standards for competent performance (ten Cate & Scheele, 2007). Put differently, if it would only be possible to anticipate what, when and how

individuals learn, it would also be possible to design assessments using predetermined accurate responses or models of performance (Delandshere, 2002; Govaerts & van der Vleuten, 2013). Nevertheless, conceiving learning as expansive (i.e. focusing upon knowledge production rather than reproduction) challenge assumptions of such predictability and uniformity in what is learned (Govaerts & van der Vleuten, 2013). Competence is more than just acquisition of specific knowledge and skills. Rather, it is about being able to generate new knowledge or skills in response to varying contexts and processes (Fraser & Greenhalgh, 2001). Therefore, learning includes discovering and acquiring things that have not been taught or acquired yet, through exchange and interactions in social networks (Engestrom & Sannino, 2010; Mennin, 2010). As opposed to traditional approaches in medical education where learning emphasises planned and formal events with well-defined and unchanging learning outcomes (Bleakley, 2010), learning is a constant process without a certain or clearly defined endpoint, and is never complete (Govaerts & van der Vleuten, 2013). In order to assess the complex and multidimensional construct of professional competence, it is important to assess learners' ability to adjust and to pliably apply and develop knowledge and skills when confronting evolving circumstances (ibid). Such ability needs to be taken into consideration as valuable and meaningful information in the appreciation of a learner's professional competence (Schuwirth & van der Vleuten, 2006).

Consequently, focusing on assessing only predetermined and specified learning outcomes would unavoidably result in oversimplification of an arbitrary phase in the process of professional development (Hager & Hodkinson, 2009; Govaerts & van der Vleuten, 2013). Assessment techniques that stay away from professional judgement in the name of objectivity may lead to an atomisation of complex skills. Hence, the content of the

assessment is trivialised (van der Vleuten, 1996). For instance, breaking down communication skills in the OSCE into its smallest behavioural components may decrease subjectivity but will not reflect the complexity of the skill (Van Thiel et al., 1992). Therefore, and although the usage of detailed yes-no checklists in OSCEs helps increase objectivity which justifies the popularity of OSCEs in clinical performance assessment (Cunnington et al., 1997), global scoring could help stop students preparing for the exam by memorising the checklists because the focus will be on skill demonstration (ibid). Behaviour that is not on the checklist, such as coherence in data gathering, can be assessed by global scoring with feedback on observed strengths and weakness (ibid). In addition, the usage of yellow and green cards in the OSCE is possible to help give more space for feedback. These two cards are used in some medical schools for noticeable strengths, green card, and weaknesses, yellow card. For instance, being rude or unprofessionally dressed would trigger a yellow card, but performance that is desirably beyond expectation would be encouraged by a green card.

Setting only closed questions in assessment methods is just like trying to shoot fish in murky water (Biggs & Tang, 2007). Therefore, it is ideal to use open-ended questions using some intended learning outcomes verbs from Bloom's revised taxonomy such as plan, produce, perform, differentiate, argue, predict, monitor, create, reflect and design (Anderson & Krathwhol, 2001). Performance assessment questions, as in the OSCE, should not underestimate the inclusion of social, cultural and ethical issues that could shape learning, learning outcomes and performance interpretations (Delandshere, 2002). This inclusion would help comprehend and assess various examples and interpretations of learning and performance in complex social settings (Govaerts & van der Vleuten, 2013).

Formative and summative assessment

Assessment can be either formative or summative depending on the function desired from the assessment. It can be formative when it guides future learning, provides reassurance, promotes reflection and shapes values. It also can be summative when it makes an overall judgement about competence, fitness to practice, or qualification for advancement to higher level of responsibilities (Epstein, 2007).

Performance of students clearly changes during learning. Competence is known to be changing and developing with time and proceed with different rates (Epstein, 2007) as deliberate practice helps practitioners gain habits of mind and practical wisdom (Ericsson, 2004) and reflection on experience (Eraut, 1994 ; Dreyfus, 2001; Epstein, 1999; Epstein, 2003; Schon, 1987). ‘‘A learner may demonstrate mastery of the required skills of physical examination and in practical procedures at the appropriate level whilst remaining deficient with regard to communication skills’’ (Harden et al., 2015, p. 37). Moreover, the influence of stress on competence might appear more clearly in less experienced individuals (Shanafelt et al., 2002; Borrell-Carrio & Epstein, 2004) regardless of the probability that all practitioners might be less competent when they are tired, distracted or annoyed (Epstein, 2007). Ongoing formative assessment and providing feedback is essential to monitor such changes and improve performance and expertise development (Norcini & Burch, 2007).

A detailed definition of formative feedback describes it as a type of assessment that functions in a formative way to the extent that evidence about the achievement of students elicited by the assessment is interpreted and used to make decisions about the next steps in teaching and learning that are likely to be better, or better founded, than the decisions

that would have been taken in the absence of such evidence (Gipps, 1994; Wiliam, 2011). Formative assessments provide benchmarks to orient the student who is approaching a relatively unstructured body of knowledge (Epstein, 2007). Such type of assessment has the ability to reinforce students' intrinsic motivation to learn and inspire them to set higher standards for themselves (Friedman Ben-David, 2000a).

Feedback is a core component of formative assessment (Sadler, 1989), central to learning, and at the heart of medical education (Branch & Paranjape, 2002). Feedback can be defined in terms of "information about how successfully something has been or is being done" (Sadler, 1989, p. 120) which can be provided by the teacher, peer or self (Hattie & Timperley, 2007). Ramaprasad (1983, p. 4) defined feedback as ''information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way". Feedback enhances students learning by informing them of their progress or lack of and, therefore, advising them regarding observed learning needs and resources available to facilitate their learning. In addition, it motivates students to engage in appropriate activities (Gipps, 1999; Shepard, 2000) because they are required to use the information in future activities (Ramaprasad, 1983).

Therefore, formative assessment is not merely intended to assign marks or grades to student performance at designated points in the curriculum; rather it is designed to be an ongoing part of the instructional process and to advocate and enhance learning (Hattie & Timperley, 2007; Shepard, 2000). The student has to fulfill three important steps in formative assessment (Sadler, 1989):

1- Possess a concept of the standard (goal or reference level).
2- Compare the actual or current level of performance with the standard.

3- Engage in appropriate action for the purpose of closing the gap.

Short-cycle formative assessments generally produce a large effect especially when they cause thinking, provide guidance on how to improve, focus on what to take forward to the next assignment, and finally are used (Wiliam, 2008). Key strategies for formative feedback include (ibid):

a) Clarifying, understanding, and sharing learning intentions.
b) Effective discussion and activities.
c) Providing feedback.
d) Collaborative learning and peer assessment.
e) Self-regulated learning, self-assessment, and motivation.

It is indeed possible to utilise the OSCE as a formative assessment tool in undergraduate medical education (Townsend et al., 2001) as it is an educational tool that provides immediate feedback (Brazeau et al., 2002; Hodder et al., 1989). ‘‘An OSCE may be administered at different times during the curriculum to assess and monitor a student’s progress and to provide personalised guidance to the student about their progress’’ (Harden et al., 2015 p. 37). An attractive feature of the OSCE is that detailed feedback can be given to the learner about areas where they have achieved the standard necessary and areas where further study is required’’ (Harden et al., 2015, p. 37). It has been clearly acknowledged that the OSCE enables assessors to identify poor performance, and appropriate remediation can then be offered (Pell et al., 2012). ‘‘The provision of feedback to a learner about their clinical competence, including their strengths and weaknesses is an important attribute of an assessment tool, particularly, but not exclusively, when the assessment is formative. The provision of feedback to students both

during and following the examination is a powerful element in the OSCE'' (Harden et al., 2015, p. 30). Increased breadth of competence, increased difficulty, increased utility and application to practice and increased proficiency are four progression dimensions identified by Harden (2007) by which a learner's progress can be assessed (Harden et al., 2015). The OSCE was appreciated by students not only as a valuable way of assessing their competence, but also as a valuable form of providing feedback (ibid). Therefore, faculty training in providing feedback is important as faculty observations of student performance may not be sufficiently accurate in both identifying and communicating errors in student performance (Norcini & Burch, 2007).

On the other hand, summative assessment is usually an assessment of learning instead of an assessment for learning (Wiliam, 2000). In the predominant view of educational assessment it is assumed that the student to be assessed has an amount of knowledge, expertise or ability, and the aim of the assessment task is to gain evidence regarding the amount of knowledge, expertise or ability (Wiley & Haertel, 1996). Consequently, summative contrasts with formative assessment in that it is concerned with reaching a decision about the achievement status of a student through conducting an assessment, usually at the end of a course or program, especially for purposes of certification. It is essentially passive and does not usually have immediate impact on learning, although it often influences decisions which may have intense educational and personal consequences for the student (Sadler, 1989). Nevertheless, "with some thoughtful planning the tutor can invariably provide quality information on individual performance. This is an area that is often overlooked because many teachers ignore the opportunity to supply feedback on summative assessment as they feel the benefits are negligible" (McAleer, 2001, p. 270).

Students tend to study what they expect to be examined on. Therefore, summative assessment may influence learning even in the absence of sufficient feedback that could drive learning (Schuwirth & van der Vleuten, 2004). Summative assessment, high-stakes examination, should not (Wiliam, 2003a):

1- Increase the link between success and self-esteem.
2- Decrease motivation for low achievers.
3- Send the message that only what is tested is important.
4- Encourage the development of shallow learning.
5- Encourage a performance orientation rather than a mastery orientation to learning.

Incorporating an OSCE in a final summative examination has become popular in many medical schools internationally (Grand'Maison et al., 1996; Boulet et al., 2009; Kim 2010). ''The OSCE can be used as a high-stakes barrier examination (summative examination) designed to certify that students have achieved the level of competence necessary to pass from one phase of the undergraduate programme to the next phase'' (Harden et al., 2015, p. 36). As previously stated, the usage of yellow and green cards in the OSCE, for noticeable strengths and weaknesses, help give more space for feedback and encouragement.

A distinction has to be made between assessments that are suitable for formative use and those that are characterised by sufficient psychometric rigour for summative use. This distinction is especially important in selecting an assessment instrument for assessing competence for high-stakes assessments such as licensing and certification examination (Epstein, 2007). OSCEs satisfy both summative and formative purposes of assessments (Harden et al., 2015).

Framework for selection of assessment methods

It is important for assessors to have a framework that can help in selecting the appropriate assessment instrument for the purpose of assessment and the required ability to be assessed. This framework needs to be based on evidence from the literature in order to avoid misapplying the selected assessment methods. For instance, the stakes are higher in summative assessment than they are in formative assessment. Therefore, assessment instruments characterised by a high degree of reliability and validity are better suited for summative assessment. Without having such a framework an assessor may select a short essay question for assessing communication skills. Communication skills require a more realistic assessment method such as the OSCE.

Historically, decisions about the selection of assessment method have rested primarily on validity and reliability (Norcini & McKinley, 2007). Recently, this has been expanded upon for the purpose of assessment in medical education. Educational effect, feasibility, and acceptability have been added (van der Vleuten & Schuwirth, 2005). The following part of this chapter explains these factors and whether the OSCE fulfils them.

Validity

This is referred to as the degree to which the inferences made about medical competence based on assessment scores are correct (Messick, 1989). ‘‘Validity describes how well one can legitimately trust the results of a test as interpreted for a specific purpose’’ (Cook & Beckman, 2006, p. 166.e8). In other words, validity determines whether an assessment method assesses what it is supposed to assess. Validity is not a property of tests, nor even of test outcomes, but a property of the inferences made on the basis of these outcomes

(Wiliam, 2000). Validity is a property of the instrument's scores and their interpretations (Messick, 1989). For example, ''an instrument originally developed for depression screening might be legitimately considered for assessing anxiety. In contrast, we would expect cardiology board examination scores to accurately assess the construct 'knowledge of cardiology', but not 'knowledge of pulmonary medicine' or 'procedural skill in contrary angiography'. Note that the instrument in these examples did not change - only the score interpretations. Because validity is a property of inferences, not instruments, validity must be established for each intended interpretation'' (Cook and Beckman, 2006, p. 166.e8). Therefore, ''one does not validate a test, but only a principle for making inferences" (Cronbach & Meehl, 1955, p. 297). This indicates that "method characteristics do not inherently determine what is being measured" (van der Vleuten, 1996, p. 51). A test item that is highly valid in one area may not be so in other areas. Consequently, validity of a test item is specific for the particular content area and for the specific purpose. A paper and pencil based test may be valid for assessing declarative knowledge but not so if the purpose is to assess communication skills.

Validity can be largely enhanced by careful operational definition of the content to be assessed (Ebel & Frisbie, 1986). Assessment is generally considered to be "a representational technique" (Hanson, 1993, p. 19) rather than a literal one. Therefore, conducting an educational assessment necessitates the ability of the result of the assessment to stand as a proxy for some wider domain (William, 2000). ''The sample tested in the examination should be representative of the learning outcome domains'' (Harden et al., 2015, p. 25). Furthermore, "competence is viewed not only as the possession of knowledge, skills, and attitudes, but rather as the ability to use these in the clinical environment to effect desired results for patients" (ten Cate et al., 2010, p. 674).

Therefore, it is ideal to consider the achieved results and impact on the environment in order to further validate the assessment instrument. Blueprinting is a process that can be used in this regard. Blueprinting, as mentioned earlier, refers to the process where the content is carefully planned against the intended learning outcomes (Dauphinee, 1994). It specifies the objectives that are to be assessed in the given assessment as well as their relative weight on the assessment. Therefore, ‘’for an examination to be valid, the content and form of the assessment needs to be aligned with the purpose of the examination and the desired learning outcomes. To be valid, the test needs to assess the learning outcome domains as defined in the curriculum and to do this through a realistic test (Harden et al., 2015, p. 25).

The OSCE was introduced to confront issues with validity and reliability. ‘‘Validity has been widely acclaimed as a feature of the OSCE and almost certainly has been an important reason for its wide adoption’’ (Harden et al., 2015, p.26). The design of OSCEs and scoring can be complex, and some challenges might arise such as deciding what to include and how to combine their scores (Hays, 2008). Harden et al. (2015) clarified that validity in the OSCE is promoted by three procedures:

1- The use of a blueprint to structure the examination. This relates what is assessed at the stations to the course learning outcomes and the body systems or other course framework.
2- The observation by the examiner of examinees in a realistic setting performing a clinical task, such as communicating with a patient, examining a patient or carrying out a procedure.

3- The assessment of both the examinee's technique and approach to the patient as well as the examinee's findings and conclusions.

Security is a critical issue in high stakes examinations. If the specific content or correct course of action is previously known, it adversely affects the outcomes and validity of the scores (Swanson et al., 1995). Securing a large pool of test material could help solve such security problems but that might have some implications for feasibility.

Reliability

Reliability is a measure of the consistency or reproducibility of a test over time, over different cases, and different examiners (Norcini et al., 1985; Wass et al., 2001). The measure of consistency over different cases (inter-case reliability) and over different raters (inter-rater reliability) have been well researched. The former measures the consistency of student's performance across different scenarios or cases while the latter measures the consistency of rating of performance by different examiners. A coefficient of 1 means that the test is flawlessly reliable as the standard deviation of the error is zero. A coefficient of zero indicates that the standard deviation of the errors is exactly the same as that of the observed test scores (i.e. the scores gained by the learners are all errors with no information about the learners at all. A test with zero reliability means that the result of the test is entirely random (Wiliam, 2001).

Inconsistency and fluctuation in performance and scoring is not uncommon. ''If a student attempts a test several times, even if no learning takes place, the student will not get the same score each time —the student might not feel very 'sharp', the marker may be more or less generous, or the handwriting might be a little bit clearer so the marker can

understand the answer. A further source of unreliability (usually the largest) concerns the particular choice of items. A test is constructed by choosing a set of items from a much bigger pool of potential items. Any particular set of items that are actually included will benefit some students (e.g. those that happen to have revised those topics recently) and not others. These fluctuations affect the quality of the information that a test gives us. For a good test, the size of fluctuations must be small in comparison with the amount of information about the individual'' (William, 2001, p. 17).

Physicians do not perform consistently from case to case (Swanson, 1987), and solutions were suggested to overcome such a reliability issue. Broad sampling across cases is crucial to assess clinical competence reliably (Wass et al., 2001). Assessing learners across a large sample of clinical cases has been a key in increasing reliability (Roberts et al., 2006). Furthermore, an appropriate assessment length has been recognised as greatly increasing assessment reliability (Swanson et al., 1995). Consequently, conducting multiple short tests is more reliable than carrying out a single long test (Wiliam, 2003b). This could explain why some traditional clinical assessments, such as long cases, had some issues with validity and reliability. The use of multiple examiners across different cases with sufficient testing time also has the potential to achieve adequate increased reliability (Norcini et al., 1985; Swanson, 1987). Using trained assessors could help reduce variation in scoring among them (Newble et al., 1980; van der Vleuten et al., 1989) as the usage of different assessors for different stations can decrease individual assessor bias (Gormley, 2011). In addition, standardised scoring rubrics helps assessors to mark learners against the same criteria thus increasing consistency of scoring between learners and assessors (Smee, 2003).

OSCEs address, to a large extent, the previous factors that could help increase reliability. The reliability of the OSCE has been comprehensively studied and well established (Walters et al., 2005; Pell et al., 2010; Boursicot et al., 2014). When designed well, OSCEs are a valid (Downing, 2003) and reliable (Boursicot, 2010) assessment method in assessing communication and clinical skills in medical and other health professions (Colliver & Swartz, 1997; Epstein, 2007; Rushforth, 2007; Sloan et al., 1996; Sloan et al., 1995). The OSCE consists of timed and different themed stations (Epstein, 2007) and the number of stations can vary from as low as 6 to as high as 40. Each domain to be examined, such as communication skills or physical examination skills, is commonly tested at several stations (Harden et al., 2015). Over the course of 3 to 4 hours, ten stations is usually the minimum number required to achieve a reliability of 0.85 to 0.90 (Epstein, 2007). Harden et al. (2015, p. 25) listed five features of the OSCE that contribute to its high reliability as a clinical assessment method:

1- Students rotate around a series of stations, where multiple samples of competence are assessed.
2- Every student is assessed on the same competencies.
3- Each student is seen by a number of trained examiners, who observe the students' performances at the stations.
4- What is tested in the examination is defined in advance, and this is reflected in the scoring sheet for each station.
5- Simulated patients (SPs) when used, present a standardised patient simulation.

The OSCE utilises simulated and real patients to allow assessors observe candidates' clinical performance. In medical education, the utilisation of standardised/simulated patients (SPs) has been widely investigated. Presenting many learners with one similar challenge helps reduce an important source of variability (Norman et al., 1985). Standardised patients can help meet specific educational goals by portraying different cases and can themselves reliably rate candidate's performance with respect to history taking and physical examinations (Epstein, 2007). Structured assessments with the use of standardised patients are usually as reliable as direct observations of encounters with real patients with no noticeable time difference (Wass, 2001). The level of reliability could be the same in both structured examinations using standardised patients and real patients when observed by the supervising assessor (Norman, 2002; van der Vleuten et al., 1991).

Nevertheless, 'performance drift' might occur when one case is played by the same simulated patient over a long period of time (McKinley & Boulet, 2004) adversely affecting the process of the assessment (Norcini & McKinley, 2007) and reliability compared to other traditional assessment format such as multiple choices questions MCQs (Clauser et al., 2002). Appropriate and good standardised patient performance helps increase reliability of the OSCE by decreasing their performance variation between leaners (Smee, 2003). As a result, it is essential to carefully choose standardised patients, train them extensively, and develop an ongoing quality assurance program (Boulet et al., 2002).

Several studies have investigated such assessment drawbacks associated with simulations, and task variability was one of the major contributors (Boulet et al., 2003; Elstein et al.,

1978; Norcini, 1999; Norcini & Boulet, 2003). Sampling broadly could help increase reliability as the performance of candidates can be patient or case specific (Norcini & McKinley, 2007). However, this increase in the number of tasks might affect cost, but it is important to note that the advantages far outweigh the drawbacks.

The OSCE uses global rating scales and checklists in order to look at candidates' performance generally and in more detail. The combination of checklists and global rating scales were found to be the most reliable assessment approach (Regher et al., 1998). Global rating scales in conjunction with checklists were employed by Hodges and McIlroy (2003) and they found that the global ratings had greater internal reliability than the checklist. Evidence has suggested that global rating scales or the combination approach used in OSCEs, global rating and checklists, can be a reliable and valid method of rating (Hodges et al., 1998).

However, and although both checklists and global scores in OSCE assessment are seen as reliable approaches (Cunnington et al., 1997), it is important to note that ''the notion of objectivity is a relative one. Even multiple-choice questions and other so-called objective tests are not as truly objective as their designers may claim'' (Harden et al., 2015, p. 1). It is important to note that, ''although the OSCE does provide a standardised and relatively objective method of evaluating a set of clinical skills in medical personnel, its use does not guarantee reliable scores and accurate decisions about medical students'' (Brannick et al., 2011, p. 1187). The reliability of an eight station OSCE was found to be low in one study (Wessel et al., 2003) recommending a larger number of stations to increase internal reliability. It has also been claimed that ''overall scores on the OSCE are often not very reliable'' (Brannick et al., 2011, p. 1181). Consensus among assessors in their judgements

would then be necessary to assess performance reliably. Such judgements need to be informed and sophisticated at a particular point in time (Govaerts & van der Vleuten, 2013). This possibility in having differences among assessors, especially in their global marking decisions, and how such differences happen is the main goal of conducting this research. The next chapter thoroughly explains what the literature says about why and how assessors make different judgements even when they observe one similar performance.

Finally, validity and reliability are closely linked. If the mark a learner gets differs radically from one occasion to another, or depends on who does the marking, validity would be affected; as there is no point in measuring one thing reliably without knowing what is being measured (Wiliam, 2001). "An assessment cannot be viewed as valid unless it is reliable" (Wass et al., 2007, p. 18). ''Reliability is a necessary, but not sufficient, component of validity (Downing, 2003; Feldt & Brennan, 1989). ''An instrument that does not yield reliable scores does not permit valid interpretations. Imagine obtaining blood pressure readings of 185/100 mm Hg, 80/40 mm Hg, and 140/70 mm Hg in 3 consecutive measurements over a 3-minute period in an otherwise stable patient. How would we interpret these results? Given the wide variation of readings, we would be unlikely to accept the average (135/70 mm Hg), nor would we rely on the first reading alone. Rather, we would probably conclude that the measurements are unreliable and seek additional information. Scores from psychometric instruments are just as susceptible to unreliability, but with one crucial distinction: it is often impractical or even impossible to obtain multiple measurements in a single individual. Thus, it is essential that ample evidence be accumulated to establish the reliability of scores before using an instrument in practice'' (Cook & Beckman, 2006, p. 166.e12).

Fidelity

Grades are supposed to represent students' attained level of achievement. One of the requirements for this property is that "all the elements that contribute to that grade must qualify as achievement, and not be something else" (Sadler, 2010, p. 727). In medical and social intervention research, fidelity refers to how faithfully the implementation of a program follows the original design (Calsyn, 2000). In regard to fidelity and assessment, for example, continuous assessment is fairer for students and more facilitative of their learning than final examinations. The issue specifically related to fidelity occurs whenever grades are accumulated across the course or program (Sadler, 2010). Assessment for learning is contingent upon judgement being based on the quality of student works, free from extraneous elements.

Furthermore, fidelity can be referred to as how faithfully a task is presented to the learner. For instance, "when the purpose of the test is limited to determining whether a student can identify the appropriate actions to take in a specific situation, such as ordering diagnostic studies, this aspect of decision making can be assessed effectively by a lower fidelity pencil-and-paper test. On the other hand, history taking or counselling tasks that require interactions with the patient are likely to require approaches of a higher fidelity, such as real or standardized-patient cases" (Norcini & Mckinley, 2007, p. 74).

The OSCE enables assessors to examine student competence at a higher level than the 'knows' or 'knows how' levels in the Miller Pyramid (Miller, 1990). It requires the learner to 'shows how' and demonstrate their competence in practice (Harden et al., 2015). For instance, the OSCE has been a common tool for assessing communication skills (Schwartzman et al., 2011) because of its ability to measure complex communication skills (Hodges et al., 1997). Therefore, ''the OSCE is what is described as

a performance test and as such is part of the movement to more authentic assessment'' (Harden et al., 2015, p. 25).

Simulation in the OSCE can help increase fidelity. It could help resemble and give the opportunity to faithfully present some tasks a doctor confronts in real practice which helps increase fidelity (Norcini & McKinley, 2007). Standardised patients allow learners to be observed as they *do* a clinical task such as interviewing or performing a physical examination while an assessor observes and rates their performance and communication skills on a standardised scale (Hodges et al., 1996). Standardised patients are credible and are usually indistinguishable from real patients (Norman et al., 1985). Students interviewing real and simulated patients did not show difference in blind ratings of empathy (Sanson-Fisher & Poole, 1980). There was no difference in the number of questions asked or the accuracy of diagnosis reached between using real or simulated patients (Norman et al., 1982). Performances with simulated patients were found to be more accurate reflections of real practice than written simulations (Rethans & van Boven, 1987). Furthermore, simulation can be less affected, than MCQs for example, by some types of security breaches because it would be difficult for any candidate to become familiar with all of the pathways through a case (Norcini & McKinley, 2007). In addition, it is more difficult for candidates to fake the correct responses. The OSCE in paediatrics was seen by students as a true measure of their essential clinical skills (Pierre et al., 2004) which provides some evidence that it is an authentic tool for assessing such skills.

Fairness

Fairness refers to the ''quality of making decisions that are not biased and are free of discrimination'' (Harden et al., 2015, p. 28). The traditional approach to clinical assessment, the 'long case', was described as unfair due to examiner bias and the fact that

the examination was conducted on a single patient (Stokes, 1974) which makes it less reliable.

Fairness is perceived by both students and examiners as a significant feature of the OSCE (Harden et al. 2015). It was seen by students as the fairest examination (Pierre et al., 2004) and as a fairer examination (McFaul et al., 1993) when compared to other traditional assessment methods. Constructive alignment and fairness are related. Following an OSCE the students commented that the examination was perceived as 'fair' because they saw that the OSCE reflected the teaching and learning programme and the stations overall addressed the learning outcomes of the course (Harden et al., 2015). The OSCE is also seen as a 'fair examination' because of the following contributions to fairness (Harden et al., 2015, p. 28):

1- All examinees have a number of tasks to perform, and these are the same for all students.
2- Examinees are assessed by a number of examiners who are briefed in advance and score the examinee's performance on an agreed checklist and ratings scale.
3- SPs give a standardised presentation and are selected by gender, age and ethnic background.
4- The rules for the OSCE are decided in advance with regard to the format, scoring approach and the standard setting procedure to be adopted.
5- What is assessed in the OSCE is closely matched with the curriculum and the expected learning outcomes.

Feasibility

Feasibility refers to the degree to which the assessment instrument selected is affordable and effective for the intended purpose (Norcini & McKinley, 2007). In other words, it refers to the degree of practicality of the assessment method, technically and economically (Harden et al., 2015). It is important to take into consideration that this could largely differ from one institution to another based on available funds and resources.

In the case of the OSCE, feasibility is identified as an important reason why this assessment method of clinical competence has been widely adopted in different contexts and situations (Harden et al., 2015). More than 1600 papers published on the OSCE validate the feasibility of the approach in a wide range of situations (Patricio et al., 2013; Harden et al., 2015). The utilisation of simulated patients is feasible, valid and moderately a reliable means of examining professional competence (Vu & Barrows, 1994). Furthermore, the usage of the OSCE is feasible even with limited resources (Harden et al., 2015). The OSCE can also be flexible which in turn increases its feasibility. The flexibility of an assessment refers to how easily it can be adapted in different situations (ibid). The OSCE approach can be adapted in different ways to suit assessors and students own needs in terms of (Harden et al., 2015, p. 27):

1- The numbers and duration of stations and the length of the examination.
2- The role of examiners and their briefing and training.
3- The role of patients, including real patients, SPs and mannequin.
4- The tasks assessed at each station and the format of the required response from examinees.

5- The use of paper or electronic recording of examiners' scores and examinees' responses.
6- The examination venue.
7- The feedback given to examinees.

Acceptability

Acceptability refers to whether medical students, faculty and patients approve the measure and the related interpretation of scores (Norcini & McKinley, 2007). It also refers to the credibility of the assessment process and results seen by stakeholders (Norcini et al., 2011).

The OSCE is internationally reported as the most preferable examination of clinical competence (Harden et al., 2015). The OSCE was perceived by educators as addressing an important issue – the assessment of clinical competence (ibid). ''Over the last 40 years, the OSCE has been widely adopted as the recommended approach to the assessment of clinical competence in different phases of education, in different specialties and in different parts of the world'' (Harden et al., 2015, p. 23). The reliability and fairness of the OSCE format over other formats of clinical assessments has helped to increase the widespread acceptability of OSCEs among students and assessors (Boursicot et al., 2014). ''Students find the OSCE acceptable because of its perceived fairness, in particular the sample of competencies assessed, the number of examiners and the transparency of the process. Teachers find the approach acceptable, in particular the authentic nature of the assessment and its validity'' (Harden et al., 2015).

Educational impact

Educational impact is now seen as one of the most valuable features of any assessment method (McDaniel et al., 2011; Roediger et al., 2011; van der Vleuten, 1996). Students focus on what they will be assessed on rather than on learning outcomes and objectives of the course (Boursicot et al., 2014). ‘‘The assessment can steer and influence the students’ learning in a desirable or undesirable way’’ (Harden et al., 2015, p. 30). The educational effect of assessment could help increase students’ motivation to do well and directs their study efforts in support of the curriculum (Norcini & McKinley, 2007). For instance, increasing learner’s knowledge requires a written assessment that will appropriately motivate learners to study from books. Similarly, increasing clinical skill would be best reinforced by a clinical assessment that helps motivate learners to interact with patients.

The OSCE has been identified as the gold standard for performance assessment (Humphrey-Murto et al., 2013; Medical Council of Canada, 2011; Sloan et al., 1995), and its impact on education has been vast (Harden et al., 2015). The OSCE can drive learning and therefore has the potential to affect how learning would take place (Boursicot, 2010). This depends on realistic assessment scenarios at the OSCE stations. If learners find it easy to differentiate between real life practice and the assessment tasks, the OSCE would not be expected to drive lifelong learning (Khan et al., 2013). In addition, if the tasks presented at OSCE stations are merely classified and driven by checklist scoring then the learners would learn to pass exams, reducing the educational impact of the OSCE (Miller, 1990; Shumway & Harden, 2003). Global scoring, as mentioned earlier, could help stop students preparing for the exam by memorising the checklists because the focus will be on skill demonstration (Cunnington et al., 1997). Behaviour that is not on the checklist,

such as coherence in data gathering, can be assessed by global scoring with feedback on observed strengths and weakness (ibid).

Conclusion

Meaningful assessment is a multifaceted process that requires the consideration of many factors. Conducting assessments properly helps to achieve the intended learning outcomes and develop a deep learning approach. The OSCE has been seen as a powerful assessment method used in assessing competence in both summative and formative ways. However, inconsistency among assessors might negatively affect reliability. This research aims to understand non-verbal behaviours that could cause such inconsistency among assessors. The next chapter will discuss in detail different theories and perspectives related to how and why inconsistency among assessors occurs.

Chapter 2-B Inconsistency among assessors

Introduction

In the last chapter, a literature review investigated the requirements and features of meaningful assessment and how the OSCE was found to be a meaningful and powerful assessment method. However, inconsistency among assessors in the OSCE might cause an issue with reliability. In this chapter, the focus is on understanding what the literature says about the factors, perspectives and different theories that could explain the possible inconsistency among assessors even when they observe one similar performance.

Assessors' marks in performance assessments can be highly variable. It has been found, in different settings, that inter-assessor disparities accounted for between 18 % (Alves de Lima et al., 2011) and 21 % (Margolis et al., 2006; Wilkinson et al., 2008) of total score inconsistency -growing to 40 % in one study (Weller et al., 2009), and considerable difference in the mean scores of assessors was highlighted (Boulet et al., 2002; Norcini et al., 1997). In another study, assessors' scores ranged from 1 to 6 on a 9 point scale while observing and assessing the same performance (Holmboe et al., 2003). Schuh et al. (2009) demonstrated that score disparities lead to unreliable pass/fail judgements by different groups of assessors. These studies clearly show that we are facing a real and significant challenge in our assessments of students' performance when we use human observation assessment tools. Inter-rater reliability, as mentioned earlier, is a main component of any assessment tool and method.

Assessment methods that use direct observation of learners have been widely employed around the world. Examples include objective structured clinical examinations (OSCEs)

(Turner & Dankoski, 2008), small-group tutorial assessments (Eva, 2001), and workplace assessments (Norcini & Burch, 2007). Human observation, represented in rater-based assessments, enables students to be observed performing complex tasks matching higher levels of competency (Fromme et al., 2009; van der Vleuten & Schuwirth, 2005). Professional competence, i.e. the use of communication, knowledge, technical skills, clinical reasoning, judgement, emotions, values and reflection, is best assessed by direct observation of the candidate interacting with a patient (Epstein & Hundert, 2002). Although assessment of students using direct observation of performance has been supported for its benefits (Durning et al., 2002; Hatala et al., 2006; Holmboe et al., 2003) and educational effectiveness (Alves de Lima et al., 2005; Holmboe et al., 2004), their utility can be limited by low inter-rater reliability (Hawkins et al., 2010; Pelgrim et al., 2011) and measurement limitations (Albanese, 1999; Lurie et al., 2009b; Williams et al., 2003).

Furthermore, taking advantage of human observation to inform assessment of its assessors and learners has faced challenges in medical education (Gingerich et al., 2011). Research identifies this struggle in that rater-based assessments generally reveal psychometric weaknesses (Albanese, 1999; Kassebaum & Eaglen, 1999; Lurie et al., 2009a; Williams et al., 2003) including measurement errors of leniency (Cacamese et al., 2007), undifferentiation (Silber et al., 2004), range restriction (Hatala & Norman, 1999), bias (van Barneveld, 2005), and unreliability (Clauser et al., 1999). It has been clearly recognised that individual examiners may vary and be inconsistent in their judgements (Burt, 1936; Roberts et al., 2010). Case-specificity and rater inconsistency have been

recognised as a source of measurement error (Clauser et al., 2008; Downing, 2005) with little understanding of how to resolve such a challenge (Gingerich et al., 2011).

Along with other features such as validity and feasibility, inter-rater reliability is always an essential feature of any assessment tool. The use of rater-based assessments in defining and assessing the competence of its learners together with difficultly in resolving the psychometric limitations of these ratings has meant that assessors are frequently blamed for the limitations of this assessment approach (Albanese, 2000; Downing, 2005; Gingerich et al., 2011; Green & Holmboe, 2010). Low inter-rater reliability is considered to be one of the biggest threats to the reproducibility of clinical ratings (Downing, 2004; Downing, 2005; Gingerich et al., 2011) and is often attributed to errors in assessors making judgements (Elliot & Hickam, 1987; Herbers et al., 1989; Noel et al., 1992). It was repeatedly found that different assessors viewing the same performance were not consistent in terms of judgements (Clauser et al., 1995; Clauser et al., 2000; Elliot & Hickam, 1987; Noel et al., 1992).

In addition, this variability, inherent in the scores, is problematic as it threatens assessment validity (Hawkins et al., 2010; Pelgrim et al., 2011) if assessors are not examining what they are supposed to examine. In one study, 19 of 20 OSCE stations each had one to eight disagreements where at least one assessor made a positive evaluative comment about a specific observable behaviour, while another assessor made a negative evaluative comment regarding the exact same behaviour (Mazor et al., 2007). Sometimes when performance assessments are subject to post hoc psychometric analysis, there is a greater amount of variance seen among the assessors (i.e. inter-rater variability) than the

learners (i.e. true performance variance) (Cook et al., 2010; Hill et al., 2009; Margolis et al., 2006). In other words, while the performance of one student is stable and consistent, different and inconsistent judgements are made by different assessors. This raises the question of why such inconsistency happens among assessors when the same performance is being observed. It is important for medical education to be able to answer such a question as this reliability issue could ultimately lead to assessments being unfair and unreliable.

The mechanisms that contribute to inconsistency in assessors' scoring remain unclear. Whether or not this issue should or can be overcome is still arguable (Clauser et al., 2008; Holmboe et al., 2011; Lurie et al., 2011). Govaerts et al. (2007) emphasise that viewing the assessor as a ''faulty instrument'' that yields inconsistent measures of an individual (hence the classical test theory notion of ''true score'' and ''error'') (Streiner & Norman, 2008, p. 170) offers a limited perspective (Yeates et al., 2012). It has been identified that assessors, as with any other individuals, could have peculiar ways of thinking and analysing tasks and situations. Marshall and Ludbrook (1972, p. 215) stated that "the judgment that an examiner makes of a candidate in the setting of the conventional test of clinical skills is an entirely personal one".

Efforts to overcome assessors' inconsistency

In order to overcome issues associated with inconsistency among assessors, researchers in medical education have attempted to readjust rating scales (Gray, 1996), forms (Silber et al., 2004), and introduce systems (Littlefield et al., 2005) to help prevent subjective prejudices and support assessor judgements during assessments (Gingerich et al., 2011). Additionally, minimising the subjectivity element through assessor training has been

emphasized due to the belief that assessors are the main problem (Green & Holmboe, 2010). However, such solutions have not had great success (Lurie et al., 2009b; Wood et al., 2006; Kogan et al., 2009). On the contrary, assessor training's limited improvement on measurement outcomes has caused some researchers to be uncertain that medical assessors are trainable (Cook et al., 2009; Gingerich et al., 2011; Williams et al., 2003). ''Some examiners are inherently consistent raters and others are less so. The former do not need training and the latter are not improved by training." (Newble et al., 1980, p. 349). Holmboe et al. (2004) and Cook et al. (2009) showed limited or no effect of assessor training in a medical education context. Neither training interventions (Cook et al., 2009; Crossingham et al., 2012; Holmboe et al., 2004) nor modifications in scale format (Cook & Beckman, 2009; Donato et al., 2008) have produced the desired degree of improvement in inter-assessor reliability (Yeates et al., 2013b). Interestingly, limiting the scale range has been identified to decrease rather than increase inter-assessor reliability (Cook & Beckman, 2009), whereas the addition of behavioural anchors has produced minor improvements (Donato et al., 2008; Yeates et al., 2013b). It could be clearly summarised that performance assessment scores are challenging, and neither modifications of scale format nor assessor training have produced the preferred enhancement in their psychometric properties (Cook et al., 2009; Green & Holmboe 2010; Holmboe et al., 2004; Holmboe et al., 2010; Lurie et al., 2009b; Williams et al., 2003; Yeates et al., 2013b).

Efforts to understand assessors' inconsistency

Domains such as psychology, social science and medical education have attempted to explain this inconsistency issue among assessors. Efforts in understanding assessors' assessment and judgement processes in order to resolve such reliability issues have been

identified and recommended (Govaerts et al., 2011; Kogan et al., 2011; Yeates et al., 2012). With such an understanding, subsequent solutions could be more valuable and evidence-based. In order to achieve that understanding, it was important to have a wider perspective and analysis of such variance among assessors which investigates different related dimensions and domains. More specifically, research was essential to investigate different proposed theories, perspectives and justifications as to why assessors make different judgements when they observe the same performance.

Research has highlighted the importance of considering assessors' social cognitive processes and equivalent implications regarding measurement of performance assessments (Gingerich, 2011). For instance, judgements are influenced by a complex and interrelated set of elements in the social setting of the assessment process, such as local norms and values, time pressure, assessment objectives and affective factors (Murphy & Cleveland, 1995; Levy & Williams, 2004; Govaerts et al., 2013). Social cognition researchers are interested in the specific cognitive processes used by people to think about the social world. They also study the processes used to make judgements and decisions (Bless et al., 2004). Seeing assessors as active information processors using judgement, reasoning, and decision-making strategies to assess learners has been highlighted (Govaerts et al., 2011). Complex interaction of impression formation has also been suggested, along with interpretation, memory recall, and judgement in assigning ratings (Mazor et al., 2007), provoking and describing possible incongruence between assessment procedures, psychometric measurement principles, and human rater capabilities (Govaerts et al., 2007; Lurie et al., 2009b; van der Vleuten & Schuwirth, 2005). Encoding, storing and retrieving information from memory, and how information is constructed and represented as knowledge has been widely investigated by social

cognition researchers. Research supports the notion that judgements rely on schema-based categorisations that are subject to social influence and lead to different judgements. For instance, experts are more sensitive to contextual cues, have quicker problem representations, make more inferences (Govaerts et al., 2011; Kogan et al., 2010).

However, it is still not very clear how such variability arises in assessors' judgements (Yeates et al., 2012). Enlightenment of the causes may help to elucidate why assessor training struggles, and suggest alternative strategies. Yeates et al. (2012) pointed out how psychological and social processes contribute to making judgements on an individual's performance had been broadly investigated within occupational psychology (De Nisi, 1984; Feldman, 1981), and their relevance to assessment within medical education was considered (Gingerich et al., 2011; Govaerts et al., 2007; Williams et al., 2003). However, very few studies in medical education have explored the processes responsible for assessors' judgements within directly-observed performance assessments (Yeates et al., 2013b). Industrial and organisational psychology researchers show that raters often possess implicit performance theories, which may vary from those specified by the organisation (Borman, 1987; Govaerts et al., 2013; Ostroff & Ilgen, 1992; Uggerslev & Sulsky, 2008). These implicit theories make it sometimes difficult to understand how a decision was made and, therefore, it might not be easy to understand why different assessors make different judgements.

Since it is vital to have a deep understanding of the underpinnings of assessor behaviours in the examination context in order to help improve our assessment procedures and techniques, Ginsburg et al. (2010) highlighted that we should consider what assessors actually observe, experience and can comment on. The process of competence assessment

is multidimensional and requires different skills and procedures. It necessitates sampling and integration of measures of performance on multiple various skills (van der Vleuten & Schuwirth, 2005). This complexity of competence assessment can also play a role in increasing or decreasing reliability as different assessors might deal with such complexity differently (Yeates et al., 2012). Since the main objective of medical education is to produce practitioners who are highly competent and capable of enhancing the health care of their patients and their communities (Frenk et al., 2010; McGaghie & Lipson, 1978), assessment in medical education tries to decide if the learner met the learning objectives, and this can be achieved by detecting predetermined observable behaviour in the learner (Gingerich et al., 2014). Thorough research has been conducted regarding the disadvantages of the current medical education system in better achieving this goal (Hodges, 2010; Irby et al., 2010), in relation to the quality of performance assessment (van der Vleuten & Schuwirth, 2005). Concerns about rater variability in performance assessments have increased, which has resulted, in a different perspective in relation to the study of the cognitive processes used by assessors.

Therefore, and in order to help increase assessment reliability, it would be helpful to look at what and how different causes might lead to inconsistent judgements among assessors. In this chapter, Gingerich et al.'s (2014) perspectives will be used as the main framework. Research from different domains and other perspectives will be synthesised, integrated and discussed within the main framework. In Gingerich et al.'s (2014) model, there are three main interrelated perspectives: (i) the assessor as trainable, (ii) the assessor as fallible, and (iii) the assessor as meaningfully idiosyncratic (Figure 4).

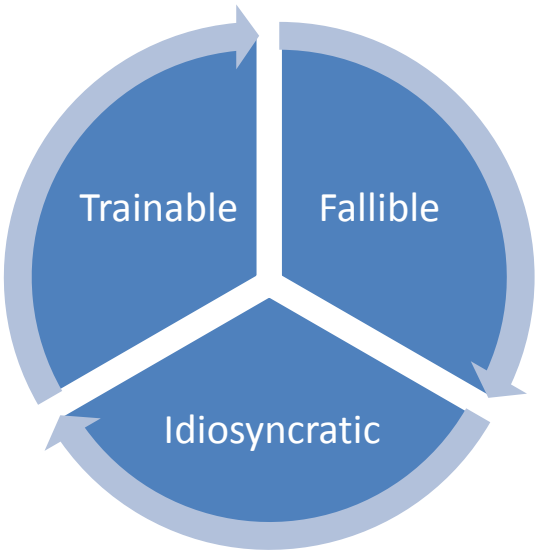

Figure 4 Gingerich et al.'s assessor model. From Gingerich et al. (2014)

Perspective 1. The assessor as trainable

This perspective refers to either the assessor applying assessment criteria incorrectly, using varied frames of reference or making unjustified inferences which lead to variance among assessors. Specifying how learning will be evaluated can be achieved through the use of assessment criteria (Ertmer & Newby, 1993; Saettler, 1990) and this requires rigorous standards for evaluating the educational objectives to ensure assessment accountability (Tyler, 1949). In this perspective, inter-assessor variability is manifested as the result of assessors either not 'knowing' or not correctly 'applying' assessment criteria. Therefore, variability in making judgements shows inaccurate information provided by assessors. In order to improve the quality of assessment information, this variability needs to be minimised (Gingerich et al., 2014). More similar responses by assessors would be expected when one student is interacting with a patient because they are observing one similar performance. However, assessors do not always succeed in appropriately using

quality metrics to assess clinical skills (ibid). The following paragraphs could throw light on the possible reasons why difficulties arise in applying metrics.

At least three key cognitive processes can adversely affect assessments when used by assessors. First, assessors use variable frames of standards against which they judge learners' performance (Kogan et al., 2011; Yeates et al., 2012; Yeates et al., 2013b). Interpreting terms such as 'satisfactory' can be very different between assessors (Kogan et al., 2009). There was significant inconsistency and uncertainty regarding how assessors translated a judgement about one learner into a numerical rating (Kogan et al., 2011). Assessors might use themselves as a frame of reference as they commonly use their own skills as comparators (Kogan et al., 2010,2011). Differences in assessors' clinical skills will ultimately lead to a clear deficiency and variance among assessors when they observe and assess students (Braddock et al., 1997; Paauw et al., 1995; Ramsey et al., 1993; Vukanovic-Criley et al., 2006). Not many assessors can explicitly apply criteria of best practice when assessing clinical performances (Hodges, 2010). Some assessors are not able to articulate what drives their assessment and can only provide a whole configuration of elements without being able to describe it as a sum of its fragments (Kogan et al., 2011). In work-based assessment, frames of reference during observation and rating enabled the comparison of the learner's performance with: (i) performance by oneself; (ii) the performance of other doctors, and (iii) a standard of performance considered to be essential for patient care (Kogan et al., 2011). One source of variability is known as criterion uncertainty. This means that assessors' criteria are uncertain, constructed differently, or influenced by recent exemplars (Yeates et al., 2013a). Assessors are expected to lack a clearly defined mental representation of the assessment criterion (Yeates et al., 2012). Assessors vary in the way they explain the elements of their

anticipation. Some assessors highlight the necessity for factual coverage; others emphasise communication or rapport building; diagnostic accuracy; or indication of developing independence (Yeates et al., 2013b).

In criterion-based assessment, assessors are also not supposed to compare between students. Assessors were found in one study to follow normative rather than criterion-referenced assessment standards (Yeates et al., 2013b). Assessors, at least implicitly, compared learners while judging their competence (Yeates et al., 2012). Instead of judging performance against fixed standards, it has been identified that assessors judge performance comparatively. Assessment should be criterion and not norm-referenced in that students are assessed according to how well they do rather than by how well they rank among their peers (Smith & Ragan, 1999; Torre et al., 2006). When assessors observe and assess students with patients, they need to be able to identify learners' 'desired' and 'undesired' behaviours. Therefore, assessors should use pre-defined criteria as many core clinical skills are associated with specific criteria by which quality care can be defined (World Health Organization, 2013).

There was noticeable disparity in assessors' perceptions of the level at which learners typically perform. These perceptions served the assessors as a general criterion, and were experientially derived, differently constructed, and frequently unclear (Yeates et al., 2013b). Hence, there were differences in comprehension and use of the assessment's criteria among assessors. Examining the influence of providing a reference point, or an anchor, on judgements made on subsequent problems, or a target, has been thoroughly investigated by psychology literature (Yeates et al., 2013a). Humans are known to be poor at judging or scaling absolute quantities; judgements are easily influenced by

contextual information (Stewart et al., 2005) through processes known as assimilation or contrast effects (Mussweiler, 2003). In one study, the preceding performances in mini-clinical examination had a noticeable effect on the score given to intermediate performances causing judgements to be prejudiced (Yeates et al., 2012). Assimilation and contrast are two opposite effects that have been reported to help comprehend what variables push assessors in one direction or the other. ‘‘In assimilation effects, a target stimulus is judged as having undue similarity to anchor stimuli. That is, scores for the target unduly reflect the anchor’’ (Yeates et al., 2013a, p. 911) For instance, assimilation in medical practice occurs when a clinician judges a patient to be severely impaired by a condition partly because a recently seen patient was severely affected (ibid). Assimilation effect is believed to occur as a result of incomplete adjustment from the original anchor (Inbar & Gilovich, 2011; Tversky & Kahneman, 1974) and because information about the anchor is selectively accessed mentally, causing similarities between the target and the anchor to be favourably perceived (Mussweiler, 2003; Yeates et al., 2013a).

Similarly, the effect of assimilation might be seen in the OSCE. The performance of one candidate can influence the score given to the next assessed candidates. The effect has been seen among both highly experienced participants at the task (Chapman & Bornstein, 1996; Northcraft & Neale, 1987) and even when participants had their own anchors (Epley & Gilovich, 2001) which makes its effect robust. Emotion might mediate assimilation effect (Yeates et al., 2013a). Sad or fearful people are more influenced with assimilation effect than angry or disgusted people (Inbar & Gilovich, 2011). It has also been seen to be inversely related to participants’ confidence when they make their judgements: participants with lower confidence displayed greater degrees of assimilation

(Jacowitz & Kahneman, 1995). However, judgemental overconfidence might be considered to be a result of representativeness bias (Tversky & Kahneman, 1974; Tweed & Ingham, 2010) (they are right less often than they think) (Yeates et al., 2013a). However, this relationship between confidence and assimilation and contrast effects was not confirmed in any other study. The contrast effect seemed to be independent of assessors' confidence in their decision (ibid).

On the other hand, the opposite of assimilation is called contrast effect. Rather than similarities, differences between target and anchor performances are overemphasised, leading to judgements that unduly vary (Yeates et al., 2013a). Contrast effects were first identified on perceptual judgements (e.g. when estimating weight or temperature) (Parducci & Perrett, 1971), by making subjective judgements of its rank position (Parducci, 1965; Wedell et al., 2005) but have afterwards been identified to occur when people judge one another (Wedell et al., 2005; Yeates et al., 2013a). Contrast effects are believed to be produced when individuals are influenced more by information in their immediate context than by information stored in their memory (Brewer & Chapman, 2003; Yeates et al., 2013a). As a result, a relatively better performance or action appears unduly good and relatively worse performance or action appears unduly poor (Wedell et al., 2005). In medical practice, this can be shown as if a clinician underestimated the severity of a patient's health condition partly because they had recently seen a patient with a worse case of the condition (Yeates et al., 2013a). Initial overall impressions play a role in whether assimilation or contrast would be produced. Initial impressions of similarity between items are inclined to produce assimilation, whereas initial impressions of difference produce contrast (Mussweiler, 2001a; Mussweiler, 2001b; Mussweiler, 2003; Yeates et al., 2013a). It has been noticed that concurrent presentations of anchor

and target stimuli are inclined to stimulate assimilation, whereas consecutive presentation stimulate contrast effects (Tanner, 2008). In the OSCE, students are observed consecutively. This may cause contrast effects. Assimilation might be considered the default, whereas tasks that include more deliberate consideration might have more tendency to contrast effects (Greifender & Bless, 2010; Mussweiler, 2003; Strack et al., 1993). Assessments of human performance findings have similarly varied. In occupational psychology, studies have presented contrast effects (Murphy et al., 1985; Becker & Villanova, 1995; Becker & Miller, 2002) whereas other studies found assimilation effects (Damisch et al., 2006). As a result, it would not be easy to generally anticipate which effect will dominate within medical education (Yeates et al., 2013a). In one study, assessors' judgements of borderline performances were vulnerable to a contrast effect (Yeates et al., 2012). Contrast effect was referred to as being where assessors who had recently observed and assessed good performances gave lower scores to borderline learners than assessors who had recently observed and assessed poor performances (ibid). Tweed and Ingham (2010) revealed that judgements on intermediate levels of performance might be particularly difficult for assessors, it is possible that contrast effects might be limited to borderline candidates (Yeates et al., 2013a).

In medical education, such as the context of the OSCE, it is expected that there is a contrast effect (ibid). This mainly happens because assessors might compare between students, and this activity has been identified to promote contrast effects (Epstein & Hundert, 2002) although comparisons between students could produce an assimilation effect (Yeates et al., 2013a). Furthermore, students in the OSCE are observed and assessed sequentially, which is also known to stimulate contrast effects (Tanner, 2008). Although it has been indicated that the contrast effect is robust, it might be possible to

mitigate it through some means. For instance, the effect could be lessened if assessors made judgements against more specific behavioural criteria (Yeates et al., 2013a). This would propose that limitations in the assessment form could increase the effect (ibid). However, this suggestion appears questionable because efforts to decrease assessor variances through improvements to assessment forms have been partially successful (Cook & Beckman, 2009; Donato et al., 2008; Landy & Farr, 1980). People meta-cognitively evaluate the suitability of their own judgements, and feel confident when this evaluation indicates that the judgement will possibly be correct (Koriat, 1993; Mitchum & Kelley, 2010). Decision confidence reflects metacognitive inferences by an individual about their adequate ability to make the judgment (Mitchum & Kelley, 2010). Tweed and Ingham's (2010) study revealed that assessors were mostly over-confident in their decisions around the threshold of adequate performance, but were under-confident at extremes of performance. In one study related to work-based assessment, it was worth noting that confidence was not largely connected to gender or frequency of conducting mini-CEX assessments (Yeates et al., 2013b).

Secondly, another source of measurement error happens when assessors do not focus on assessing observable behaviours, but make inferences during direct observation (Govaerts et al., 2013; Kogan et al., 2011). Inferences about trainees' knowledge, skill and attitudes are made by many assessors (Govaerts et al., 2011; Kogan et al., 2011). Such inferences are not recognised by assessors, and therefore they do not validate them for accuracy (Kogan et al., 2011). These invalidated inferences risk misrepresenting the accurate assessment of the learner as assessor's inferences cannot be observed and measured; and this ultimately might lead to low inter-assessor reliability (Gingerich et al., 2014).

Thirdly, some assessors might modify their assessment judgements for unrelated reasons. Although it is not true of all assessors, some assessors may inflate assessments in order to be perceived as popular and likable (Kogan et al., 2011). In addition, some assessors tend to be a bit more lenient to avoid defending their assessments and conversations with institutional leaders (Cleland et al., 2008; Dudek et al., 2005; Kogan et al., 2011). Harasym et al. (2008) investigated assessment approaches in undergraduate family medicine objective structured clinical examinations (OSCEs) and found that eliminating hawkish (stringent) and doveish (lenient) influences changed the outcome for around 11% of learners. In one study including 2000 assessors (McManus et al., 2013), around 2% of them were statistically significant hawks and 2% significant doves.

Perspective 2. The assessor as fallible

Fundamental limitations in human cognition, related to human memory and processing capacity, can cause low inter-assessor reliability, meaning that immediate context randomly influences assessors. It has been recognised that learners and professionals are not constantly performing at their best, and that performance might differ from day-to-day or even within the same day (Govaerts & van der Vleuten, 2013). Performance can lack temporal stability, particularly in very complex tasks (Fisher, 2008; Sturman et al., 2005). This could easily apply to assessors which leads to variances and inconsistency in their judgements. Reasons could be motivational, physiological such as fatigue, or any other unstable cause affecting individual performance, such as mood or emotional experiences (Beal et al., 2005), and environmental factors (i.e. opportunities and constraints in the context setting, even in experts and talented performers) (Govaerts & van der Vleuten, 2013). Vulnerability to environmental constraints differed across individuals and job complexity, signifying that performance is determined by the

interaction between individual, task and environment (Govaerts & van der Vleuten, 2013; Stewart & Nandkeolyar, 2007). True intra-individual performance disparity could result from changes in the individual (e.g. due to motivation, fatigue, changing levels of competence) as well as changes in the job environment and context (Govaerts & van der Vleuten, 2013; Sturman et al., 2005). Intensive training of assessors might not make a noticeable difference (Landy & Farr, 1980) and many different researchers challenge seeing the assessment process as a 'precise analytical machine' (Gingerich et al., 2014). In this perspective, low assessor-reliability happens as a result of fundamental limitations in human cognition. Human judgement is flawed and can always be influenced (ibid). Cognitive and social psychology affirm that assessors cannot simply and perfectly observe and capture performances (Ilgen et al., 1993) as human memory and processing capacity are imperfect (Baddeley, 1994). Information can simply be lost very quickly (van Merrienboer & Sweller, 2010). Within-person disparity in performance is significant and can be as large as between-person variances (Fisher & Noble, 2004; Deadrick et al., 1997; Stewart & Nandkeolyar, 2007).

This could be applied to both learners and assessors. Performance assessments can be unacceptably biased, suffering from halo and leniency effects, and intra- and inter-assessor reliability of performance assessments are often found to be substandard (Cook et al., 2010; Kreiter & Ferguson, 2001; van Barneveld, 2005). Consequently, pure objective observation of performance does not exist (Gingerich et al., 2014). Maintaining and comparing information long enough to give scores and feedback necessitate that humans interfere with what they observe. Such cognitive processes can be the source of different biases (Macrae & Bodenhausen, 2001) and the foundation of problems with

judgement-based assessments. Our tendency to categorise people results in the notion of a typical person (Gingerich et al., 2011) which cause a risk our judgements being biased (Tversky & Kahneman, 1974). Rather than carefully processing and dealing with all available information, humans tend to compare key features of one person with those of a 'typical' example (Gingerich et al., 2014).

Beliefs about others carry important messages. Trusting or avoiding someone, for example, may happen as a consequence of what belief someone has about other people. Beliefs about others play an important role in our different life activities. This importance increases when it is related to making judgements and assessing students. Assessors are responsible for making fair and accurate judgements. Consequently, issues that can affect this accuracy of making judgements such as beliefs about others and making global impressions need to be understood and investigated in order to overcome possible flaws associated with making these judgements. Global impression formation, categorisation or stereotyping cause what is known as the halo effect, which is commonly attributed to assessor error when they perceive others (Govaerts et al., 2013). The study of the beliefs that people have about others is usually referred to as 'Person Perception' (Kenny, 1994). In every interpersonal perception, there is a perceiver, target and a trait. Using traits was a dominant way to describe others (Fiske & Cox, 1979) especially in the absence of a verbal trait label (Winter & Uleman, 1984). In this model, the perceiver (medical assessor) rates the target (medical student) on a given trait or behaviour. Generally, perceptions can be classified into three main types: self- perception, meta-perception, and other-perception (Kenny, 1994). Self-perception refers to how someone sees him or herself. Meta-perception is the process of 'mind reading' where a person may attempt to discern how he or she is seen by others. Other perception, or the perception of the other,

is simply how you see other people and how others see you. The latter is what this chapter intends to focus on and discuss although the three types can affect each other. Person perception is reciprocal or two sided which means that people simultaneously perceive each other, whereas object perception is one sided. The second difference is that in person perception people usually spend some time attempting to read other people's minds. These two points may play a significant role in person perception due to the possible impact they have. Before proceeding, the next paragraph succinctly explains the possible effect and relationship between the three types of perceptions, self- perception, meta-perception, and other-perception.

The three types of perceptions are interrelated. Nine fundamental questions of interpersonal perception were raised in order to manifest the relationship among these perceptions (Kenny, 1988; Laing et al., 1966; Malloy & Albright, 1990; McLeod & Chaffee, 1973; Scheff, 1967; Tagiuri et al., 1958). The nine questions can be described as follows:

1- *Assimilation* means whether a perceiver sees two targets as similar. There is evidence that people tend to see other people as more similar than they really are (Kenny, 1994).
2- *Consensus* refers to whether two perceivers agree when they judge a target. However, it should be clear that even if two perceivers agree, it does not necessarily mean that they are accurate. There is evidence to show that increasing acquaintance does not result in greater consensus.

3- *Uniqueness* concerns the extent to which a perceiver views a target idiosyncratically. It is a dominant component in the perception of others or other-perception. There are three sources that uniqueness effects can be attributed to:
   a- Two perceivers use different information to judge a target.
   b- Two perceivers attach different meanings to the same observed behaviour.
   c- Perceivers may apply non-behavioural information such as his or her liking of the target in the ratings.
4- *Reciprocity* refers to whether a perceiver and a target see each other similarly. There is little evidence for reciprocity (Kenny, 1994).
5- *Target accuracy* refers to the validity or accuracy of other-perception.
6- *Assumed reciprocity* concerns the extent to which a perceiver thinks that a target sees him or her as he or she sees that target. In other words, do people think that others see them as they see others?
7- *Meta-accuracy* refers to the extent to which people are good mind readers. In other words, do people know what others think of them?
8- *Assumed similarity* concerns how a person sees others and how he or she sees himself or herself. In other words, does a person think that people are similar to him or her?
9- *Self-other agreement* concerns the correspondence between how others see a person and how that person sees himself or herself.

In the context of the OSCE, the perceivers would be the assessors and the targets would be the candidates. The following table summarises the nine questions classifying them into groups according to their relationship with the three types of perception.

Table 5 Nine fundamental questions of interpersonal perception

The degree of similarity or lack thereof between two other-perceptions	Accuracy or validity of other-perception	The degree of similarity between other-perception and meta-perception	The relationship between self-perception and other-perception
1- Assimilation 2- Consensus 3- Uniqueness 4- Reciprocity	5- Target accuracy	6- Assumed reciprocity 7- Meta-accuracy	8- Assumed similarity 9- Self-other agreement

It has been increasingly emphasised that assessors are to be understood as 'social perceivers' providing 'motivated social judgments' when assessing performance (Murphy & Cleveland, 1995; Klimoski & Donahue, 2001; Levy & Williams, 2004; Govaerts et al., 2013). When perceivers are interacting with others, they are expected to be concerned with self-presentation and do not have the cognitive capacity to process all of the target individual's information (Gilbert & Jones, 1986; Gilbert et al., 1987). They are seen as active information processors who confront cognitive tasks of gathering, interpreting, integrating and retrieving information for the process of judgement and decision making that takes place within a dynamic and complex social setting (DeNisi, 1996; Donahue 2001; Govaerts et al., 2013; Klimoski & Donahue, 2001; McGaghie et al., 2009). Information processing by assessors is affected by their definition and understanding of effective performance, personal goals, interactions with the learner and others, as well as by other elements in the social context of the assessment process (Govaerts et al., 2007; Govaerts et al., 2013; Murphy et al., 2004; Uggerslev & Sulsky, 2008).

Forming an impression of an individual mirrors an integration of all the information available to characterise that individual. (Anderson, 1968; Asch, 1946). Information integration refers to assessors explaining the valence of their comments in their own

unique narrative terms, usually leading to global impressions formation. While assessors make judgements, they explain and probably mentally represent the valence of those judgements in unique narrative terms. These unique narrative and global judgements, in turn, are converted into the assessment scale to produce scores for each individual domain (Yeates et al., 2013b), which ultimately leads to inconsistency and low inter-rater reliability. Perceivers usually construct impressions from factual information, inferences, and evaluative reactions regarding the target person (Hamilton et al., 1989; Gingerich et al., 2011). Impressions help organise information into a structure of knowledge about that person (Lingle et al., 1979) in order to be able to interact with him or her (Leyens & Fiske, 1994). Within psychology literature, the process of perceiving other people, known as impression formation, is commonly referred to as 'categorization task', though different cognitive processes are thought to be enacted. (Macrae & Bodenhausen, 2000; Fiske, 1993; Gingerich et al., 2011). Categorical judgements are formed about learners as part of forming impressions. However, impression formation researchers have not underestimated the importance of investigating the idiosyncrasy of assessors (Kenny, 2004).

Interestingly, the descriptions made by a single assessor about several candidates have been found to be more similar than the descriptions made by several assessors about a single candidate (Bourne, 1977). It has been found that social judgements are idiosyncratic and fallible under certain situations and conditions (Nisbett & Ross, 1980). For instance, assessors' mood and emotions can have an influence at the time of the judgment (Forgas, 1994). In addition, a ratee that reminds the rater of a significant other can affect the ratee to be perceived to share similar characteristics with that significant other (Anderson & Cole, 1990). ''Impressions are subject to variables and contextual

factors beyond the ratee himself or herself'' (Gingerich et al., 2011, p. 52). Impressions have frequently been regarded as personal to the assessor and effortlessly biased by numerous factors (Skowronski & Carlston, 1989; Williams et al., 2003). It has been well established that different assessors will often form different impressions of the same learner even when given the exact same information (Kenny, 1994; Park et al., 1994). Assessor's unique way of translating techniques can cause errors in assessment systems that require ordinal or interval ratings when assessors form categorical judgements. However, and regardless of the possibility of having idiosyncratic categorisation, assessors tend to constantly make one of a few possible interpretations of each learner (Gingerich et al., 2011).

The social cognition literature highlights three themes that summarise the differing notions of categorisation as used in forming impressions of other people. These themes are as follows: (i) the conceptualisation of impression formation as the construction of Person Models, (ii) impression formation as a nominal categorisation process, and (iii) impression formation as a dimensionally based categorisation process (Gingerich et al., 2011).

A- Impression formation as the construction of Person Models

Impression formation has been conceived as a procedure whereby perceivers generate *person models* of other people, explaining what the person is like and why (Park, 1986; Park et al., 1994). The Person Model is based on the building of stories, as required, to describe specific individuals (Mohr & Kenny, 2006; Park et al., 1994). It has been suggested that the process of forming person models was one of storytelling or narrative development (Mohr & Kenny, 2006). Perceivers go beyond listing personality traits that

explain a target individual by integrating underlying explanations as to why the person behaves the way they do or possesses the particular traits (Mohr & Kenny, 2006). According to Fiske (1993, p. 170), "faced with surprising combinations for which they do not possess ready-made structures, people create brief stories that provide enabling and temporal links among otherwise puzzling bits of information." It has been identified that the 'person model' shares several features with theories that emphasise the use of social categories as a means to interpret and integrate information about a ratee (Skowronski & Carlston, 1989; Fiske, 1993). The suggested reason for having multiple stories for each ratee relies on different combinations and prioritisation of the pieces of information by assessors (Park et al., 1994). This variance ultimately affects assessment reliability and is frequently described as noise resulting from the idiosyncrasy of the rater (Mohr & Kenny, 2006). Perceivers do not systematically allocate targets to person models, nor does agreement exist regarding the traits of individuals. Perceivers seem to latch onto a model that organises several elements of the perception of individuals. Social interactions provide considerably more information to perceivers than provided in a "static stimulus display" (McArthur & Baron, 1983). For example, judging someone's cooperativeness is unviable without observing the person engaging in social interaction (Mohr & Kenny, 2006). However, it is worth noting that the relatively short time of an OSCE station interaction might not provide assessors with all the required information.

Several research studies attempting to document agreement in personality judgements have instead found that these judgements are more frequently unique than similar, even when perceivers are presented with the same information about a person (Kenny, 1994; Park & Judd, 1989). Person models can assist in describing why there seems to be little agreement in personality judgements; perceivers generating multiple models of a person

would also view them differently regarding trait and affect ratings. However, it is not clear how perceivers reach their distinct views (Mohr & Kenny, 2006). It has been found that perceivers who share similar models agree on personality trait inferences, which could illuminate the process by which perceivers integrate information about someone into a coherent impression (Asch, 1946; Mohr & Kenny, 2006). Another assumption of person model formulation is that perceivers spontaneously shape different information about a target into an integrated impression (Park et al., 1994). Park et al. (1994) claimed that each perceiver, when forming a model, would attend to a certain characteristic and construct an impression around that central notion. One central piece of information about a person is expected to be affective judgment (Park et al., 1997). Other research supports the impact affective reactions have on impressions of people (Schneider et al., 1979).

Gender-related behaviours are another possible type of information that could be used to organise target information into a person model, specifically masculinity–femininity (Brewer, 1988; Brewer & Harasty Feinstein, 1999; Fiske et al., 1999; Fiske & Neuberg, 1990). In one study, it was found that person models were valid across multiple groups of perceivers, and that the models, from presidential candidates to co-workers, varied in terms of masculinity–femininity and affect (Mohr & Kenny, 2006). Stereotype researchers claim that gender is a salient cue in the environment (Chiu et al., 1998; Rothbart & John, 1993) and the most prominent of categories upon which people base their impressions (Stangor et al., 1992). Nevertheless, not all perceivers would apply the gender category similarly when judging an individual (Stangor, 1988). It is worth noting that, and based on the extent to which perceivers attended to it, masculine and feminine models were potential for both men and women, contingent on their behaviour (Mohr & Kenny, 2006). Although target gender was found related to person model formulation, it

was not the case in all the ways expected. ‘‘Men were not systematically assigned more masculine models than women were, nor were women systematically assigned more feminine models than men. Rather, the masculinity and femininity of each person model was potentially related more to the sex-typing of behaviors’’ (Mohr & Kenny, 2006, p. 348).

Exemplar-based models, whereby target individuals could call to mind other people with comparable core features (Andersen & Cole, 1990), would be another possible potential for person models formulation (Mohr & Kenny, 2006). However, some researchers argued that one difficulty with the exemplar-based explanation is the very limited number of person models that emerged (Park et al., 1994). If the basis for impression formation was relying on similarity to significant others, the number of person models would not be very limited. Impressions will frequently be quite consistent across raters regardless of the expectation of raters being idiosyncratic when forming impressions (Gingerich et al., 2011). Although an infinite number of person models could possibly be generated about one particular person (Mohr & Kenny, 2006), and probably perceivers are permitted to choose among the possibilities (Wittenbrink et al., 1998), Park et al. (1994) suggested that there are typically two or three reasonable models. In another study, although assessors had different conceptions of competence in multi-source feedback, their conceptions were grouped into four diverse constructs of competence (Thammasitboon et al., 2008). Therefore, person perception is found to be idiosyncratic, yet consensual. Descriptions of a ratee, in two studies, written by raters based on their impressions could be grouped into three representative stories (or “Person Models”) about that particular rate (Park et al., 1994; Mohr & Kenny, 2006) with one story being more common than the others (Gingerich et al., 2011). It is important to notice that the same three models are not

relevant to every other ratee, however (ibid). Consequently, even though judgements can be idiosyncratic, they are not infinitely so.

It was demonstrated by Park et al. (1994) that disagreement arises from the different procedures in which perceivers shape information that they receive about a target, rather than from differential judgements of individual acts. Perceivers look for consistencies in behaviour and annotate over situational disparity when forming impressions (Mohr & Kenny, 2006). Park et al. (1994) gathered written descriptions of 25 target people's behaviour in each of five different situations and found that perceivers imposed larger consistency in their impressions when viewing behaviours that they knew were from a single target than when they were unaware. In the latter circumstance, situational information carried noticeably more weight in determining the perception. In the OSCE, it is possible to see assessors who know the students for a long time as their class students, and it is also possible to have assessors who have never met the students before. This might influence an assessor's judgement by their preexisting perception of a candidate's performance, attitude and skills.

B- Impression formation as a nominal categorisation process

The focus in this suggested etiology of assessor error is not on the particular construction of narratives around someone's behaviour; rather, it is more about assessors' tendencies to co-locate or gather candidates into preexisting schemas (Gingerich et al., 2011). As a level of measurement, the nominal scale "classifies objects into categories based on some characteristic of the object" (Hurlbert, 2006, p.15). Assessors are not scaling the behaviours differently but, rather, they are assigning learners to different nominal categories. Unlike the first assumption, this process exists in the long-term memory and is

applied when activated (Macrae & Bodenhausen, 2000). It has been believed that assessors use categories in applying preexisting knowledge to comprehend incoming information about a person (Gingerich et al., 2011). Research has manifested that the process of assessing and judging others can be illustrated as a categorisation task that proceeds through a combination of automatic and deliberate cognition to allocate individuals to categories, in part based on similarity (De Nisi, 1984; Feldman, 1981; Gingerich et al., 2011; Govaerts et al., 2007; Williams et al., 2003).

Categorisation can carry some dangers as in overgeneralisation (i.e. stereotyping), but it may also have some benefits (Macrae & Bodenhausen, 2000). Categorisation avoids the cognitive resources used to monitor a ratee's category-consistent behaviour. Instead, the assessor is only required to note any category inconsistent behaviours (Macrae et al., 1994). It also allows the assessor to go further beyond the available information to infer other anticipated details consistent with typical category members (Sherman et al., 1998) and then decide how to behave when interacting with them (Fiske, 1993).

Category-based knowledge, consistent with the Person Model theories of impression formation, is believed to provide some justifications for why a learner might demonstrate particular behaviours in a given situation (Gingerich et al., 2011), and it explains what a group of people are like and why (Wittenbrink et al., 1998). Social categorisations of any individual are assumed to be flexible because any one can be categorised in several ways (Stangor et al., 1992). With regard to how controllable category activation is, some researchers claim that it is automatic and not controllable (Bargh & Ferguson, 2000). Stereotypes are expected to be (i) unintentional and (ii) occur without perceivers' awareness which represent two of the criteria commonly related to an automatic mental

process (Bargh, 1989). Other researchers argue that perceivers' awareness is not always absent, and this makes it either conditionally automatic (Monteith et al., 1998; Gilbert & Hixon, 1991) or consciously controllable (Blair & Banaji, 1996).

B-1 Stereotype

Although it saves a lot of mental effort, categorising and comparing learners means that assessors tend to not use important information, thus risking judgements becoming prejudiced. This type of prejudice is well explained by the literature on stereotypes. Impressions of people influenced by their membership of a group rather than their individual features, stereotypes (Gingerich et al., 2014), can misrepresent which features individuals focus on (Bodenhausen & Wyer, 1985), the decisions they reach (Bodenhausen, 1988) and their recall of what occurs (Dijksterhuis & van Knippenberg, 1995). People may not be aware when their judgement of someone is influenced by their stereotypical beliefs (MacRae et al., 2002) either due to influence on cognition (Nisbett & Wilson, 1977) or behaviour (Bargh & Chartrand, 1999). Stereotypes have frequently been characterised by social psychology as energy-saving strategies that assist the important cognitive function of simplifying information processing and response generation (Allport, 1954). Individuation, in its many pretences, is a rather time consuming and effortful matter (Brewer, 1988). Stereotyping, in contrast, relies only on the implementation of some rather basic skills: most particularly, the ability to assign people to meaningful social categories (Hamilton, 1979). The concept of stereotypes as simplifying mental devices is not new. Perceivers appear at best reluctant in individuating others unless a series of critical cognitive and motivational criteria (Macrae et al., 1994), such as spare attentional resources, outcome dependency and accountability have been satisfied (Erber & Fiske, 1984). Stereotypes serve as energy-saving or resource-

preserving mental devices (Allport, 1954), and Lippman (1922) claimed that reality is excessively complex for any individual to represent precisely. Stereotypes, therefore, help to simplify perception, judgement, and action. Information processors when challenged by limitations would necessitate compromises and shortcuts. Fiske and Neuberge (1990, p. 14) commented, ‘‘we are exposed to so much information that we must in some manner simplify our social environment. For reasons of cognitive economy, we categorize others as members of particular groups- groups about which we often have a great deal of generalized, or stereotypic, knowledge.’’ Fiske (1989, p. 253) justified, ‘‘stereotypers categorise because it requires too much mental effort to individuate’’. Stereotypical thinking is an omnipresent feature of everyday life (Macrae et al., 1994). Gilbert and Hixon (1991) characterised stereotypes as tools residing in a metaphorical mental toolbox. ‘‘Although there are clearly cases in which those who stereotype do pay a penalty (e.g., failing to hire the best job applicant because of gender stereotypes), the act of stereotyping may typically produce errors that are more costly to others than to the perceiver him- or herself’’ (Macrae et al., 1994, p. 44). This can be clearly seen if it happens in assessing students.

Intentionally trying to adjust social judgements or suppress categorical thinking can have a negative influence on impressions (Wegner, 1994). Assessors who tried to avoid the use of stereotypes demonstrated more stereotypic thinking in subsequent judgements (Macrae et al., 1994) and more stereotyped memories of the candidate (Sherman et al., 1997). Therefore, trying to avoid categorising or stereotyping people might not be entirely possible and may not succeed in producing better judgements. On the contrary, it might make the problem worse (Macrae et al., 1994) which means that simple training will not necessarily overcome the problem (Gingerich et al., 2014).

Answers to a question like ‘‘how friendly is someone?’’ might be shaped and coloured by understanding of the implications of someone's behaviour, profession, age, gender, ethnicity, interpersonal relations, personality traits, physical appearance, abilities, goals, family background, or any other information about them could be considered related (Kunda & Thagard, 1996). Both stereotypes and individuating information are processed simultaneously, and jointly affect impressions of others (ibid). The previous diverse selections of information that can colour impressions of others are classified by social psychologists into two major types—‘stereotypes and individuating information’ (Brewer, 1988; Fiske & Neuberg, 1990; Locksley et al., 1982).

Individuating information, such as traits or behaviours, and social stereotypes jointly influence impressions of individuals. ‘‘Stereotypes refer to membership in social categories such as gender, race, age, or profession that are thought to be associated with certain traits and behaviours. Individuating information refers to anything else known about the individual behaviour (e.g., hit someone), personality (e.g., introverted), family circumstances (e.g., has two brothers), etc.’’ (Kunda & Thagard, 1996, p. 284). Stereotypes might have a larger influence on impressions when observed before, rather than after, individuating information has been observed (Bodenhausen, 1988). Stereotypes could dominate impressions when they are observed before individuating information, but individuating information could dominate impressions in the same way when it is observed first, assuming that both types of information possess equal status when they are observed simultaneously (Kunda & Thagard, 1996).

The influence of stereotypes and individuating information on how someone forms impressions of others could be clarified using some phenomena described by Kunda and Thagard (1996).

Phenomenon 1: Stereotypes colour the meaning of behaviour

One similar ambiguous behaviour displayed by two different candidates, who belong to different social categories, could be interpreted differently. For instance, it was found that a shove was viewed as more violent when performed by one person from one social category than by another person who belongs to another social category (Duncan, 1976; Sagar & Schofield, 1980). Likewise, this would also raise the idea that two assessors from different backgrounds or social categories could have different interpretations or acceptance of a behavior displayed by one candidate.

Phenomenon 2: Stereotypes color the meaning of traits

One similar trait can suggest different behaviours when applied to candidates of different social groups. For instance, Kunda et al. (1995) identified that perceivers rated lawyers and construction workers as about equally aggressive. However, they held different anticipations about their probable aggressiveness - related behaviours: lawyers were expected to argue and complain while construction workers could punch and yell. Similarly, some OSCE assessors, for example, might hold different anticipations about male and female candidates' likely politeness or kindness related behaviours.

Phenomenon 3: Stereotypes in the absence of individuating information colour impressions

Assessors might sometimes be unwilling to apply a stereotype to a candidate when they don't have enough information because they feel it is inappropriate to do so (Darley & Gross, 1983). However, there is evidence that stereotypes do colour one's beliefs when no additional individuating information is available (Kunda & Thagard, 1996). For instance, an individual who was described only as a 'night person' was viewed as more unpredictable than an individual who was described only as a 'day person' (Locksley et al., 1982). Similarly, if an individual is described only by a male name, he was viewed as more assertive than an individual who was described only by a female name (Locksley et al., 1980; Rasinski et al., 1985). Therefore, gender-related stereotypes, for instance, might be activated by some OSCE assessors whenever there is a lack of some necessary information.

Phenomenon 4: Stereotypes can provoke contrast effects on trait ratings.

For instance, individuals from one race were typically viewed as less academically competent than individuals from another race. However, a member of the first race with strong academic credentials was viewed as even more competent than individuals from the second group with comparable credentials (Jackson et al., 1993; Jussim, et al., 1987; Linvflle & Jones, 1980). This could be directly equated to medicine as there are studies which show that non-white students often don't do as well in performance examinations as white students as will be discussed later.

Phenomenon 5: Stereotypes affect impressions in the presence of truly irrelevant information

It was recognised that stereotypes might influence impression formation by negligible information. For instance, minimal biographical information such as name, age and address were found to influence impressions (Yzerbyt et al., 1994).

Phenomenon 6: Non-diagnostic but pseudo-relevant information can eliminate or dilute the effects of stereotypes.

It was found that pseudo-relevant information could reduce the influences of stereotypes on judgements of the candidate (Kunda & Thagard, 1996). For instance, a description of a learner that involved information about parents' occupations, which were quite unremarkable and unrelated to the dimension of self-control, served to eliminate the influences of some stereotypes on judgements of the student's self-control and other stereotypic traits (Locksley et al., 1982).

Whenever resources are limited, stereotypes have a greater impact on impressions, either because perceivers are not at their optimal time of day (Bodenhausen, 1990), because they are cognitively busy (Gilbert & Hixon, 1991; Pendry & Macrae, 1994), because they are happy or angry (Bodenhausen et al., 1994), or are under time pressure (Dijksterhuis & van Knippenberg, 1995; Kruglanski & Freund, 1983). Furthermore, circadian rhythms (Bodenhausen, 1990), pre-existing levels of prejudice (Kunda & Spencer, 2003), and individual cognitive preferences (Crawford & Skowronski, 1998) play a role on how stereotypes might affect judgements. The influence of stereotypes might decrease when perceivers are required to pay more attention and effort on the judgement (Kunda & Thagard, 1996), either because it is complex (Bodenhausen & Lichtenstein, 1987) or

because they anticipate that they will be accountable for their decisions (Bodenhausen et al., 1994).

Motivation (Fiske & Neuberg, 1990; Klein & Kunda, 1992; Kunda et al., 1990), emotions and affect (Bodenhausen et al., 1994; Esses et al., 1993; Jussim et al., 1995) are believed to have an influence on stereotypes and judgements processes. Whenever individuals lost the ability or motivation to think more deeply about members of stereotyped groups, stereotypes were activated (Bodenhausen, 1990, 1993). It is improbable that all of the possible stereotypes that can characterise a given person will be activated at the same time. In most cases, only a subset of these will be activated (Kunda & Thagard, 1996).

Some medical researchers consider categorical thinking (Gingerich et al., 2011), cognitive load (Tavares & Eva, 2013; Wood, 2013) or first impressions (Wood, 2014) as potentials that can influence assessors' judgements. Examiners in objective structured clinical examinations (OSCEs) are expected to experience mental workload that is higher than what occurs in other routine clinical work (Byrne et al., 2014). Making detailed checklists might not always help in improving objectivity as the possibility of cognitive load increases (Tavares & Eva, 2013). The context plays an important role in determining which stereotypes are activated (Macrae et al., 1995). ''Assessor behaviours are framed within the context in which assessment takes place'' (Govaerts & van der Vleuten, 2013, p.1169). Although stereotypes and their influence on assessors have been well established in social psychology, the influence of stereotypes in medical education and assessment judgements is unclear.

B-2 Bias

Bias is another threat that has been significantly discussed because of the negative impact it possesses on assessment reliability. Some assessors declared an awareness of different biases or favouritism when assessing in real life (Yeates et al., 2013b). Learners' scores differ according to their ability which leads to 'true variance' in their scores. However, 'error variance' might result from a range of other sources such as variable case difficulty, variable differences in behaviour between patients, and from differential marking behaviour among and within assessors due to different biases (Denney et al., 2013). When assessors observe and judge the performance of candidates' face-to-face, they are able to identify candidates' gender and ethnicity and possibly infer where their initial degree was obtained. The potential for prejudiced or unfair treatment might arise from systematic bias of parallel subgroups of assessors, in both written and performance tests (ibid). McManus et al. (2008) described unexplained differential performance in medical examinations among different subgroups of learners as a phenomenon that challenges the discipline of medical education seriously. For instance, Denney et al. (2013 p. 718-719) referred to some studies that have addressed this issue:

''Dewhurst et al. (2007) found sex and ethnicity differences among UK graduates (UKGs) in the intercollegiate MRCP Part 2 Clinical Examination (PACES) in 2003–2004 (male candidates failing at 1.5 times the rate of female candidates, and black and minority ethnic (BME) candidates failing at 1.7 times the rate of the white candidates); the latest (2012) published statistics show sex differences (UKG male candidates failing at 1.3 times the rate of UKG female candidates), ethnicity differences (BME UKG candidates failing at 1.3 times the rate of white UKG candidates), and differences with regard to source of primary medical degree (IMGs failing at 3.1 times the rate of UKGs). The clinical examination of the Royal College of Paediatrics and Child Health (the MRCPCH) reports differences (2010– 2011) with regard to sex ('pass-rate is between 15% and 25% higher for females as compared to males') and with regard to source of degree, with IMGs failing at approximately 1.9 times the rate of UKGs. The clinical examination of the Royal College of Psychiatrists (the MRCPsych CASC) reports (2008–2010) sex differences (UKG male candidates failing at 1.5 times the rate of UKG female candidates), ethnicity differences (BME UKG candidates failing at 2.4 times the rate of white UKG candidates), and differences with regard to source of primary medical degree (IMGs failing at 4.7 times the rate of UKGs on first attempt). A major review and meta-

analysis by Woolf et al. (2011) of performance by ethnic group in UK examinations showed consistent underperformance by BME medical students and postgraduates, across UK specialties.''

Some research indicated that there was no association between learner and examiner gender and a minor but highly significant interaction of learner and examiner ethnicity on stations assessing communication skills and ethics (Dewhurst et al., 2007). The reduced academic achievement of students from some ethnic minorities can indicate unconscious bias (van der Bergh et al., 2010). In another study, there was no sex bias and possible ethnic bias in only one case (McManus et al., 2013), while a different study indicates that assessors show no general inclination to favour their own kind (Denney et al., 2013). Although very little is known about such effects in objective structured clinical examination (OSCEs), bias is still possible especially bias associated with assessor demographics.

C- Impression formation as dimensionally based categorisations

In this model, judgements are made on dimensional scales. Two dimensions constitute the basis of social judgements that can account for the bulk of inconsistency in impression formation. One refers to socially desirable or undesirable traits that can have a direct impact on others such as being friendly or honest, and negative traits such as cold or deceitful. The second dimension refers to traits that tend to more directly influence the individual's success such as being intelligent or ambitious, and negative traits such as being indecisive or inefficient (Abele & Wojciszke, 2007; Fiske et al., 2007; Gingerich et al., 2011). Park et al. (1994, 1997) and Wittenbrink et al. (1998) claimed that affect would be an important dimension of person models. It guides the preliminary interpretation of behavioural information by re-evaluating and adjusting the level of liking, which in turn enhances the affective evaluation of the person (Mohr & Kenny, 2006). Target and

perceiver discrepancies may be higher in cross-cultural judgements, due to dependence on consensual stereotyping (Kenny, 2004).

Different labels have been given to these dimensions, likely attributable, in part, to differing domains having been studied: warmth/competence (Fiske et al., 2007; Judd et al., 2005), communion/ agency (Abele & Wojciszke, 2007; Ybarra et al., 2008), social/intellectual (Rosenberg et al., 1968), morality/ competence (Wojciszke, 2005), and social desirability/ social utility (Beauvois & Dubois, 2009) with a common overlap of traits and behaviours identified by researchers from different domains (Abele et al., 2008; Abele & Wojciszke, 2007; Beauvois & Dubois, 2009; Fiske et al., 2007; Judd et al., 2005). For instance, researchers have clarified that the stereotyped groups can be categorised into each cluster based on assessor judgements of warmth/competence dimensions and that each cluster is connected with emotional and behavioural responses in the assessor (Fiske et al., 2002). More specifically for example, Gingerich et al. (2011, p. 4) stated:

''*In North America, groups judged high on warmth and competence, such as the middle class, invoke the emotions of pride and admiration and lead to behaviors of wanting to help and associate with them. Groups judged low on warmth and high on competence, such as the stereotypically gluttonous rich, elicit envy and willingness to associate but also to attack under certain conditions. Groups judged high on warmth and low on competence, including stereotypes for the elderly and disabled, elicit pity and willingness to help but also to avoid. Low judgments of both warmth and competence, including stereotypes for the homeless and drug-addicted, invoke the emotions of disgust and contempt and lead to behaviors of wanting to attack and to avoid*.''

These two orthogonal dimensions are dichotomised into high- versus low-value judgements. When the two dimensions are crossed, therefore, the result is four possible groupings, and it has been suggested that individuals and groups are categorised in one of these four groups (Cuddy et al., 2007). The finding that two dimensions would account

for a difference in impression formation is interesting because two dimensions have also been found to lie beneath rater-based assessments in medical education (Gingerich et al., 2011; Nasca et al., 2002; Ramsey et al., 1993; Silber et al., 2004). Assessment forms intended to assess clinical competence usually recognise two underlying factors notwithstanding the number of items or the number of dimensions involved on the form. Of the two factors that explain the majority of difference in judgment formation, one is inclined to refer to knowledge and the other to interpersonal skills. The knowledge dimension appears to be equivalent to the competence dimension in social judgements, and the interpersonal skills dimension appears to be equivalent to the warmth dimension. Therefore, medical assessors may be using cognitive processes, described above using the example of stereotyped groups in North America, to categorise learners into one of the four clusters with consequent emotions and reactions (Gingerich et al., 2011).

In medical education, the majority of rater-based assessments require assessors to rate using a predefined list of performances and competencies. These assessment dimensions might not resemble the categorisations that result from assessors' innate cognitive processes, and they might not be totally applicable to all candidate categorisations. Errors, therefore, might originate from assessment systems asking assessors to perform judgement tasks that are different from cognitive processes used by humans when performing judgements (Gingerich et al., 2011). For instance, if assessors are developing nominal judgements but assessment forms require ordinal or interval ratings, it would be interesting to know how that categorical judgement could be translated into a rating scale value (ibid).

Perspective 3. The assessor as meaningfully idiosyncratic

In the first two perspectives, differences in assessors' judgements are manifested as being problematic when making assessment judgements, and efforts have been made to find solutions to overcome the issue of low inter-assessor reliability. In the third perspective, the view of differences among assessors is completely different. Experienced assessors are more capable of making sense of highly complex scenarios which suggests that assessor variance may characterise legitimate experience-based interpretations (Gingerich et al., 2014). Assessors' variability that is based on relevant but different, and maybe conflicting, interpretations can provide thorough and meaningful assessment.

It has been established that learning, competence (as inferred from performance) and performance interpretations are to be realised as inherently contextualised, and can only be understood as a socially located interpretive act (Govaerts & van der Vleuten, 2013). Learning is conceived as inherently situated, collaborative, transformational and expansive (i.e. paying attention to knowledge production rather than reproduction) (ibid). This challenges expectations of predictability and uniformity in what is learned and what is to be learned. Ginsburg et al. (2012) stated that junior doctors' approaches to professional dilemmas relied on some factors such as individual patient characteristics, the doctor's affective response and relationship with the patient. From a socio-cultural standpoint, performance needs to be socially constructed and determined by an individual's perception of and interaction with situational features of the task at hand (Govaerts & van der Vleuten, 2013).

Applying this framework to the assessment of performance: performance is always conceived and constructed according to the perspectives and values of an individual

assessor, influenced by their experiences and the social structures in the assessment task and its context (Gipps, 1999; Govaerts & van der Vleuten, 2013). Socio-cultural approaches to assessment proclaim that assessors are no longer seen as passive measurement instruments, rather as active information processors who interpret and create their own personal reality of the assessment context (Govaerts & van der Vleuten, 2013). Delandshere and Petrosky (1994, p. 16) declared: ‘‘Judges’ values, experiences, and interests are what makes them capable of interpreting complex performances, but it will never be possible to eliminate those attributes that make them different, even with extensive training and “calibration”. Variances in an assessor’s interpretation and scoring of performance could be equally valid (Landy & Farr, 1980) and meaningful (Lance et al., 2008). Assessors’ behaviours and assessment outcomes could be affected by a broad range of context factors, such as interpersonal relationships, emotional and cultural factors (Ferris et al., 2008; Tziner et al., 2005;). Assessors, in work-based assessment, engage in complex and unpredictable tasks, influenced by time pressures and conflicting as well as ill-defined goals (Levy & Williams, 2004; Murphy & Cleveland, 1995).

Variance attributable to idiosyncrasy of assessors or context-specific variation might, from a psychometric measurement standpoint, be considered to contribute to measurement error. However, and according to situated cognitive and socio-cultural theories, context is an active and interchangeable detail that is not separate from a learner’s performance. Context both enables and constrains the learner’s ability to perform any anticipated or required skills (Durning et al., 2010; Durning & Artino, 2011; Richter Lagha et al., 2012) because it involves and covers all the interactions taking place in that unique context (Durning et al., 2010; Durning & Artino, 2011; Engestrom & Sannino, 2010; Hager, 2011). Real learner performance differences attributable to context

or case specificity are recognised to play an important and large role in the complexities of assessor-based assessment (Eva, 2003; Gingerich et al., 2011). Paying attention only to predefined learning outcomes makes assessment oversimplification of an arbitrary stage in the process of professional development (Hager & Hodkinson, 2009; Govaerts & van der Vleuten, 2013). Social, cultural and ethical issues that colour and construct learning, learning outcomes and performance interpretations need to be considered in assessment objectives (Delandshere, 2002). This suggests that the aim of assessment is not to 'objectively' and 'accurately' measure learning or learning outcomes, but to comprehend what, how and why learners are learning by comprehending and explaining context (Moss et al., 2006). Interpretivist, social-constructivist and socio-cultural approaches (Delandshere, 2002; Johnston, 2004; Koch & DeLuca, 2012; Moss, 1996; Moss et al., 2006) suggest that performance assessments are seen as social constructions or interpretations, rather than absolute, objective truths (Johnston, 2004); single true score or objective rating of performance does not exist (Gingerich et al., 2014). Rather, truth is all about consensus among assessors who need to reach judgements on performances that are informed and sophisticated at a particular point in time (ibid). According to concepts of learning and performance based in sociocultural theory, assessment should not just focus on learning outcomes, but also on the processes underlying learning, performance and performance interpretations in dynamic and complex contexts (Govaerts & van der Vleuten, 2013). Assessors' thinking could be masked when they rely on a set of scores because it is going to be about quantification. Therefore, interpretive assessment could be trustworthy (Delandshere & Petrosky, 1998) as it requires more descriptions and interpretations. Challenges in modern health care practices and education would not be easily confronted with exclusive focus on psychometric discourse (Hodges, 2013;

Schuwirth & Ash, 2013). However, both psychometric-based and constructivist-interpretivist assessment approaches confirm that inferences about professional competence need to be credible and defensible (Kane, 2008).

As a result, context will play a large role in shaping and revealing a learner's competence (Ginsburg et al., 2012; Kuper et al., 2007). Competence is socially constructed and needs to be perceived by others (Delandshere & Petrosky, 1998; Hodges, 2006; Lingard, 2009). Some key constructs that require assessment are not directly observable (Pangaro & ten Cate, 2013). For example, constructs such as patient-centeredness and professionalism (Delandshere & Petrosky, 1998; Kuper et al., 2007), or responsibility and praise and blame (Malle & Pearce, 2001; Read et al., 1990; Reeder et al., 2002; Weiner, 1995) need to be inferred from observable demonstrations.

It is essential to distinguish between inaccuracy and idiosyncrasy. It has been assumed that when we sample enough to ensure adequate reliability, we are able to provide learners with accurate feedback (Yeates et al., 2013b). This meaningful idiosyncrasy has raised a concern about how to triangulate between several perspectives to make a broad picture of a learner and deliver valuable feedback about their performance (ibid). Govaerts et al. (2007) have discussed this idea and affirm the approach to programmatic assessment recommended by van der Vleuten and Schuwirth (2005). Govaerts et al. (2007) suggest a theoretical understanding of performance assessment based on a constructivist, social psychological perspective. This affirms that social and cognitive factors interrelate and interact to produce idiosyncratic individual judgements on performance (i.e., inconsistency that can be attributed to meaningful variances in the perceptions of assessors) (Yeates et al., 2013b). This affirms that since some

inconsistency might arise from the well-recognised cultural or other biases that raise concerns about the reliability and validity of an assessment practice, some could arise simply from individual peculiarities in approach- such as unique ways in which the task is understood or judged (ibid).

Regardless of being possibly meaningful, idiosyncrasy within assessor cognition can lead to low inter-assessor reliability. Research in various domains claims that idiosyncratic assessor effects account for large differences in performance assessments, ranging from 29 % to over 50 % (Hoffman et al., 2010; Scullen et al., 2000; Viswesvaran et al., 1996). Rater idiosyncrasy levels are considerable and are not always related to rater expertise (Govaerts et al., 2013). The focus of this chapter is to better understand the sources of such meaningful idiosyncrasy in order to help utilise the advantages and avoid the drawbacks that can result from assessors being idiosyncratic.

Perceivers tend to utilise pre-existing knowledge structures, or 'schemas' when they perceive others and make judgements. Schemas are illustrated as adaptive mechanisms that permit people to competently process information, especially in circumstances where information is partial, vague or where there are situational constraints such as time pressure or work load (Govaerts et al., 2013). Three types of schemas are used by most people in social perception: role schema, event schema, and person schema (Pennington, 2000). A role schema is the sets of behaviours anticipated of an individual in a certain social position (e.g. a dentist, general practitioner, family physician). Event schemas explain what we generally anticipate from other people's behaviours in certain social situations, related to the anticipated sequence of events in such a situation (e.g. a job interview or performance appraisal interview). Person schemas mirror the inferences we

make about an individual on the basis of incomplete available information, through verbal and non-verbal interactional cues in their behaviour (Govaerts et al., 2013). Person schemas might contain anticipated patterns of behaviour, personality traits and other inferences about an individual's knowledge base or social category (for example, excellent or poor performer) (ibid). When people observe others, these three schemas together are used interactively to guide the focus of their attention, what they recall and how they handle information in formulating impressions and making judgements (Pennington, 2000). In OSCEs, assessors are expected to be aware of the role and event schemas with some slight differences with regard to how much is expected from each student. Therefore, it is anticipated that person schemas will play a great role in judgement differences.

Findings from industrial and organisational psychology are supported by recent research in medical education (Govaerts et al., 2013; Kogan et al., 2011). Govaerts et al. (2013) explored the usage of performance theories by experts. Their findings indicated that assessors, when observing and assessing performances, used general as well as task-specific performance theory and person schemas to make decisions about performance effectiveness (Govaerts & van der Vleuten, 2013). Personal theories and performance constructs were the basis for the process of making and justifying judgements by assessors (Govaerts et al., 2013). It is proposed that assessors in work settings develop personal constructs or 'theories' of efficient job performance overall (Ginsburg et al., 2010). These 'performance theories' are very analogous to role schemas in that they contain clusters of effective behaviours in relation to any number of performance dimensions considered relevant to that specific job (Govaerts et al., 2013). Performance theories progress and advance through professional experience, socialisation and training.

Therefore, the content of performance theories is expected to differ from one assessor to another, causing variant levels of assessor idiosyncrasy (Uggerslev & Sulsky, 2008). Consequently, assessor idiosyncrasy in the interpretation of task performance could be a result of differing individual experiences, beliefs and professional values (Govaerts & van der Vleuten, 2013). It has been indicated that the specific cluster of behaviours related to effectual performance might vary from one task to another, depending on the setting and specific characteristics and structures of the task (e.g. physicians utilise diverse communication strategies depending on situational demands) (Govaerts et al., 2013; Veldhuijzen et al., 2007). During task performance observation, task- or situation-specific cues may activate the usage of task- or event-specific schemas to assess performance, particularly in more experienced assessors (Govaerts et al., 2013). As a result, the notion, underlying psychometric assessment theory, of the reality of a single true score, is challenged by the assessment that is framed in socio-cultural constructivist theories. Rather, training, socialisation and task experience play a large role in how assessors construct and reconstruct their own performance theories and conceptualisations of competence (Govaerts & van der Vleuten, 2013). Assessment outcomes, therefore can be influenced by all kinds of schemas described above: assessor's personal performance theory ('role schema'), normative anticipations of task-specific behaviours (task- (event-) specific schema) and inferences about the learner (person schema) (Cardy et al., 1987; Borman, 1987). Table 6 (Govaerts et al., 2013, p. 381) gives a clear picture of the three different schemas discussed above.

Table 6 Verbal protocol coding structures (p.381)

Performance theory: performance dimensions and sub dimensions
1. Think and act like a general practitioner
2. Doctor-patient relationship
2.1. Atmosphere
2.2. Balanced patient-centeredness
2.2.1. Develop and establish rapport
2.2.2. Demonstrate appropriate confidence
2.2.3. Demonstrate empathy/empathic behaviour, appropriate for problem
2.2.4. Open approach
2.2.5. Facilitating shared mind 1 = identifying reasons for consultation; exploring patient's perspective
2.2.6. Facilitating shared mind 2 = explain rationale for questions, examinations; explain process; share own thinking
2.2.7. Facilitating shared mind 3 = collaborative decision making
3. Handling (bio)medical aspects (disease)
3.1. History
3.2. Physical examination
3.3. Diagnosis/differential diagnosis
3.4. Patient management plan
4. Structuring of the consultation and time management

Task- (event-)specific schema
1. Identification of case-specific cues
1.1. Specific aspects of the patient's problem/clinical presentation (e.g. this type of eczema poses very serious social problems to the patient)
1.2. Specific aspects of the patient's behaviours (verbal as well as non-verbal; e.g. this patient is very talkative)
1.3. Setting/context of the medical consultation (GP's office versus outpatient clinic)
2. Trainee behaviours (effective or ineffective) within performance domain X, explicitly related to case-specific cues
3. Effects of trainee behaviour on patient behaviour/doctor-patient relationship (positive or negative)

Person schema
1. Inferences regarding
1.1. Knowledge base
1.2. Personality traits (e.g. he is a very nice guy)
1.3. Disposition (e.g. this trainee has a clinical method of working; finds it difficult to just lean back and listen to what patients are saying)
1.4. Intention (e.g. he seems to be focused on the biomedical aspect of the patient's problem)
1.5. Category (e.g. he is an authoritarian doctor; he will become an excellent doctor)
2. Phase of training (frame of reference for making judgments)''

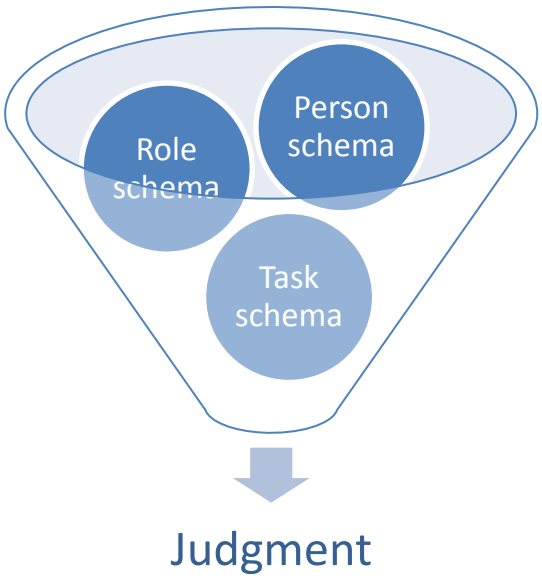

Figure 5 The combination and influence of the three models on judgments

It is worth noting that the usage of different schemas (Govaerts et al., 2013) appear to be inconsistent with some research on work-based assessment demonstrating that assessors possess a one- or two-dimensional notion of professional competence (cognitive/clinical and humanistic/(psycho)social) and are therefore incapable of discriminating between different competencies or dimensions (Archer et al., 2010; Cook et al., 2010; Pulito et al., 2007).

It has been increasingly suggested that assessor expertise and diagnostic expertise are related in different domains (Berendonk et al., 2013; Govaerts et al., 2011,2013). Experienced clinicians do not use the detailed checklists novices use regarding signs and symptoms. Instead, experienced clinicians use rapid, automatic pattern recognition to form diagnostic impressions and group sets of information into meaningful patterns, enabling fast and accurate diagnostic reasoning (Gingerich et al., 2014; Gruppen & Frohna, 2002). Therefore, experts rely on the identification and interpretation of related contextual cues. In addition, experts are able to identify irregularities that interrupt anticipations, identify what actions have already done, and form expectations of actions that are expected to happen based on the current situation (Chi et al., 1981; Klein, 2009; Norman et al., 1985). Similarly, experienced assessors are able to identify situation-specific cues in the assessment process, relate case-specific cues to case-specific performance requirements and performance assessment and form comprehensive interpretations of performance (Govaerts et al., 2011, 2013). As expertise progresses through immersion within a specific context (Webster-Wright, 2009), assessors develop a unique cognitive filter that result from exposure to different contexts, unique experiences, and different models of general and typical performance (Govaerts et al., 2011,2013). Assessors will utilize their past experiences and understandings of their social, cultural

and contextual surroundings to understand and interpret actions taking place in that unique context. Consequently, different interpretations of complex performances are expected (Gipps, 1999; Kuper et al., 2007) because of subjectivity elements in identifying and interpreting relevant cues. As mentioned earlier, Delandshere and Petrosky (1994) declared that assessors use their judgements, values, experience and interest in order to interpret complex performances, and those attributes that make them different will never be possible to be eliminated, even with extensive training and "calibration". Experts and inexpert assessors differ in how they use different schemas, and information processing appears to be influenced by differences in assessor expertise (Govaerts et al., 2013). It was found that experienced assessors used task-specific performance schemas more considerably compared to inexpert assessors, which indicates more differentiated performance schemas in expert assessors (Yeates et al., 2013b). Expert assessors were found to develop problem representations more quickly, were more sensitive to contextual cues, and made more inferences, compared to non-experts (ibid). Consequently, and although score disparity might still exist among experts, expert assessors seem to have more detailed assessment schemata than non-experts (ibid). Govaerts et al. (2007) proposed that score disparity could be regarded as a form of idiosyncrasy rather than merely as error. Therefore, it could be suggested that expert assessors are more able to provide a thorough and detailed feedback to candidates, and this can be well utilised in the OSCE.

Generally, people make social inferences spontaneously (Macrae & Bodenhausen 2001; Uleman et al., 2008), and raters' person schemas might guide selective attention in following assessments and influence the interpretation of future information. As a result, assessors' idiosyncratic interpretive processes might cause clear differences in person

perception (Mohr & Kenny, 2006). How different assessors form person schemas in WBA contexts might consequently be one of the main factors underlying differences in assessment outcomes (Govaerts et al., 2013). Correct connections and correspondence between different performance dimensions can be high, and observed halo effects might be considered, at least partially, as 'true halo' rather than as a consequence of assessor lack of skill or automatic categorisation of learner performance (ibid). Evidence, as mentioned earlier, has supported that social-cognitive processes that underlie judgements, such as stereotyping, are very flexible and adaptive to the assessor's social goals, motivations, emotional state and relationships with others (Smith & Semin, 2007). Similarly, initiation and application of mental representations or knowledge structures, such as person schemas, previously thought to be subconscious and automatic, are influenced by the social context where assessors make their judgements (Govaerts et al., 2013). It has also been found that dimensions are of variable degrees of importance (Ginsburg et al., 2010). Assessors value or pay different degrees of attention to different aspects of the performances while observing students (Yeates et al., 2013b). Assessors' experience can shape and colour stated assessment criteria, which can result in focusing on different aspects of performance. This results in different definitions among assessors of what determines quality (Kogan et al., 2011; Yeates et al., 2013b). The aspects of the performances that assessors considered useful varied, regardless of observing the same performance. Assessors' attentional focus and possibly the weight they allocate to different aspects of performance varies (Yeates et al., 2013b).

Conclusion

Different factors were found to play a role in decreasing inter-rater reliability in assessments that use direct observation of learners such as the OSCE. Researchers

attempted to readjust rating scales and forms and minimise subjectivity through assessor training. However, such solutions were not very successful. As a result, many theories and perspectives were proposed to explain why inconsistency among assessors exists even when they observe one particular encounter. Research from psychology, social sciences and medical education all together manifested different causes and justifications for such a reliability issue.

Assessors as trainable, fallible, and meaningfully idiosyncratic were the main framework in this chapter. Under each perspective of these three perspectives, different elements and concepts were synthesised and discussed. The 'assessor as trainable' perspective refers to the incorrect application of assessment criteria, using different frames of references, or making unjustified inferences. This perspective explains why a norm-referenced assessment might be applied when examining candidates by the OSCE even though it is required to follow a criterion-referenced assessment as discussed in the previous chapter. The 'assessor as fallible' perspective refers to the fundamental limitations in human cognition, such as imperfect memory, that could cause issues with assessment reliability. It also describes the processes of impression formation and the effects of stereotype and bias on reliability. The last discussed perspective was the 'assessor as meaningfully idiosyncratic'. In this perspective, score disparity can be seen as meaningful rather than merely as error. Making sense of highly complex scenarios varies from one assessor to another which suggests that assessor variance may characterise legitimate experience-based interpretations of both verbal and nonverbal consultation skills. Therefore, such disparity could be used in the OSCE as a meaningful feedback to candidates either in a summative or formative way as mentioned earlier in the previous chapter. The previous three perspectives could raise a question about the complexity of competence assessment

in medical education. More specifically, a question could be raised about how assessors' global marking can be influenced by different sources. For instance, it would be interesting to know how non-verbal behaviours influence assessors' global marking when they observe and assess medical students.

A question for research

After completing the two literature chapters, a question for research was raised and asked. Since inter-rater reliability can be influenced by different sources, the question was posed about the influence of non-verbal behaviours on assessors' judgements and the complexity of competence assessment. How such behaviours influence assessors' decisions is what this research aims to answer and investigate. In this project, the aim was to conduct research that could help improve inter-rater reliability among OSCE assessors by understanding how non-verbal behaviours can influence assessors' global marking when they observe and assess undergraduate medical students. The research question is stated as: *'How non-verbal behaviour influences assessors' global marking when they observe and assess undergraduate medical students using objective structured clinical examinations'*. The next chapter will describe how this research was conducted and what research philosophy was used.

Chapter 3 Methodology

The previous two chapters reviewed the literature, and then a question was asked to help understand how inter-rater reliability can be influenced by non-verbal behaviours. In this project, the aim is to conduct research that could help better understand some issues related to inter-rater reliability among OSCE assessors by understanding how nonverbal behaviours can influence assessors' global marking when they observe and assess undergraduate medical students.

Research question

'How non-verbal behaviour influences assessors' global marking when they observe and assess undergraduate medical students using objective structured clinical examinations.'

Research philosophy and approach

A well-defined research strategy that uses an unbiased and robust framework can help me provide unbiased and robust outcomes (Wilmot, 2005). Bunniss and Kelly (2010, p. 358) stated that ''the quality of research is defined by the integrity and transparency of the research philosophy and methods, rather than the superiority of any one paradigm''. The conceptual framework within this study would dictate whether the researcher is conscious of it or not, what I choose to do, and how I interpret my outcomes and results (Bordage, 2009) in order to fulfill the required integrity and transparency.

The research paradigm is a main part of any research project that needs to be clearly described and justified (Reeves et al., 2008). Describing only the techniques and tools I used for data collection and analysis would not be adequate as the tools themselves are not the essence of the paradigm (Lingard, 2007). Paradigms can be defined as sets of

beliefs and practices which could regulate research within fields of study, and shared by communities of researchers. Every paradigm is characterised by ontological, epistemological and methodological differences in how to conceptualise and conduct research and contribute to knowledge construction (Weaver & Olson, 2006). Within medical education research, there are four major paradigms currently in use: positivism, post-positivism, interpretivism, and critical theory (Bunniss & Kelly, 2010; Denzin & Lincoln, 2000; Guba, 1990; Guba & Lincoln, 1994; Schwandt, 1990). Each research paradigm could generate valuable information (Karlsson & Tham, 2006), but it is important, especially in medical education, to articulate the research assumptions about ontology (nature of reality), epistemology (nature of knowledge) and methodology (nature of research) and the related research methods and tools in order to follow the most relevant and effective research paradigm (Denzin & Lincoln, 2000; Guba, 1990; Schwandt, 1990).

Methodology is ‘‘a philosophical stance or world view that underlies and informs a style of research’’ (Sapsford & Jupp, 2006, p. 175). Although the term ‘methodology’ is usually used to indicate an applied approach to a specific issue, it might not always be used within medical education journals to explain research methodology and the related ontological and epistemological perspectives (Bunniss & Kelly, 2010). Therefore, and in order to decide which research paradigm is most suitable and relevant to answer my research question, the following four questions need to be clearly answered (Allen et al., 1986; Denzin & Lincoln, 2000; Guba & Lincoln, 1982; Schwandt, 1990; Weaver & Olson, 2006):

1- *What is the nature of reality (ontology)?*

There is no single ultimate reality this research aims to find. Non-verbal behaviours that could influence OSCE assessors can be subjective and change from one place to another. Therefore, the outcomes do not necessarily need to be generalised because a different reality might be constructed differently in a different context.

2- *What is the nature of knowledge (epistemology)?*

There is no single perfect way of knowing reality because knowledge obtained in this research is subjective and there are many interpretations of reality. Nevertheless, listening to what my research participants would say is important for data collection. Such data, obtained from listening to interviewees, could help elaborate on and explain hidden and complex issues related to making decisions and judgements when examining undergraduate medical students using the OSCE.

3- *What is the nature of the approach to research (methodology)?*

Diverse interpretations were gathered in this research using a modified grounded theory approach. There was more focus on comprehending and using inductive, rather than deductive, reasoning. Meaning was constructed, through the analysis of my data, in the researcher-interviewee interaction in the natural environment.

4- *What techniques could be used to gather such required information (methods)?*

More reliance on qualitative methods, such as interviews or focus groups, could be used in this research in order to interpret different views and study complex, unstable and non-linear social change (Berwick, 2008). However, this research used 1:1 interviews and not focus groups because the latter might hinder some OSCE assessors from revealing some

sensitive topics or information. In addition, some assessors might dominate the discussion.

Consequently, and after answering the previous four questions, the most relevant framework and paradigm that could be applied to conduct this research would be 'Interpretivism' (Bunniss & Kelly, 2010). This paradigm is capable of studying and exploring diverse and contextually dependent issues which is essential within medical education research (ibid). Interpretative perspectives, with other perspectives such as phenomenology and hermeneutic perspectives, are embraced within a broader framework, called constructivism (Guba & Lincoln, 1994). ''Constructivism is the view that knowledge, and therefore all meaning, is not discovered but socially constructed. Meaning is not created but constructed out of the world that is already there, and objects in that world. The world and its objects may have no intrinsic meaning, but they are partners in the generation of meaning'' (Illing, 2010, p. 288). This research tried to interpret and construct knowledge about how nonverbal behaviour influences assessors' global marking in the context of the OSCE at the university of Leeds. The interpretation and construction of knowledge in this research was based on the views and subjective beliefs and experience the interviewed OSCE assessors had. Therefore, the interpretation and construction of knowledge might be different if this research took place in a different context.

This project is classified as qualitative, and not quantitative, research. The former aims to gather an in-depth understanding of human behaviours, and investigates the questions *why* and *how* instead of *what*, *where* or *when (*Denzin & Lincoln, 2005). Qualitative research was chosen in this research because it provides understanding of the world as seen through the eyes of the individuals being studied (Wilmot, 2005). My research

question is more related to social and human experiences, and such questions are better answered by qualitative research (Denzin & Lincoln, 2005). Qualitative research has been used in medical education after its value was proven in research from the social sciences and humanities, from different disciplines such as anthropology, sociology, education and history (Lingard & Kennedy, 2010). Quantitative research, on the other hand, which usually begins with a hypothesis, and the research tests that particular hypothesis (Denzin & Lincoln, 2005) was not used in this project.

Modified grounded theory

Qualitative analysis does not end with categorising and building themes. Rather, it is essential to interpret what has been found. This interpretation process can be conducted using different qualitative research approaches. One is known as 'grounded theory' which was developed by Glaser and Strauss in the 1960s to focus on generating theory instead of testing it (Glaser & Strauss, 1967). Unlike positivists who choose an existing theoretical framework, and then collect data to manifest whether the theory applies to the phenomenon under study or not, qualitative researchers in grounded theory construct theory through the analysis of data (Faggiolani, 2011; Martin & Turner, 1986).

However, it is possible to adapt grounded theory to suit studies being undertaken as there is no one way of undertaking grounded theory studies (Bulawa, 2014). The initial approach by Glaser and Strauss was never intended to be inflexible (Glaser & Strauss, 1967). Reviewing literature has enriched this research with theories and perspectives that could help link my findings together. Strauss and Corbin (1990) emphasised that reviewing the literature in grounded theory studies is important for qualitative researchers to detect relevant categories and understand their relationships. In addition, and as a way

of stimulating theoretical sensitivity, reviewing the literature helps in ‘‘providing concepts and relationships that are checked out against actual data’’ (Strauss & Corbin, 1990, p. 50).

Furthermore, some of the questions I asked my interviewees were shaped by my understanding of related literature discussed by other researchers. For instance, some questions about using different frames of reference were shaped by my understanding of Gingerich et al.’s (2014) perspectives. Strauss and Corbin (1990) argue that as a qualitative researcher I can use literature in obtaining a range of questions (Appendix 3) to be asked to my interviewees and validating the accuracy of my findings. I started by collecting a small set of data, ‘‘guided by the initial research questions’’ (Punch, 2001, p. 167), to be analysed, before another set of data was collected with the guidance of the emerging themes and categories coming from the initial analysis. Therefore, the questions I asked my interviewees were subject to adjustments during the period of data collection and analysis based on discovered ideas, themes and categories. It is important to state that data interpretation was interpreted and constructed based on the participants’ views and experiences. Strauss and Corbin (1998) emphasised the necessity of maintaining an analytic distance from what is already known in order to be impartial and accurate in data interpretation.

Recruiting and sampling

The process of selecting a random sample is ideal, well defined and rigorous for quantitative research. However, the same technique could not be used for this current qualitative research. The reason lies in the aim of the study; ‘‘studying a random sample provides the best opportunity to generalize the results to the population but is not the most

effective way of developing an understanding of complex issues relating to human behaviour'' (Marshall, 1996, p. 523). As mentioned earlier, the aim of this qualitative research was not to generalise findings, but to understand complex issues related to how assessor's judgement can be influenced by non-verbal behaviours in the context of OSCE. This requires interpreting and constructing knowledge obtained from a specific group of participants who could provide the required data. Therefore, the process of sampling here is different from what is applied in quantitative research.

Different approaches to selecting a sample for qualitative research have been suggested such as convenience sampling or judgement sampling (Marshall, 1996). Convenience sampling involves the selection of the most accessible subjects, which makes it the least rigorous approach. This project followed the second approach, judgement sampling, which is also known as purposeful sampling. In order to answer my research question, assessors who had been involved in assessing medical students using OSCEs needed to be interviewed.

An invitation was electronically sent via an e-mail to a database of OSCE assessors who had taken part in assessing undergraduate medical students in the University of Leeds, England. Generally in qualitative research, and in order to describe a phenomenon of interest and answer a research question like mine, researchers usually collect data and simultaneously construct theory from the collected data. When no new or relevant pieces of information are obtained, saturation is reached (Given, 2008). The theory in this research project appeared robust with no unexplained perspectives or phenomena after I interviewed 18 assessors. In addition, in qualitative research, smaller but focused samples are usually used instead of large samples (Denzin & Lincoln, 2005). As a result, and in reaching theoretical saturation, 18 participants were enough to recruit to take part in this

study, 11 males and 7 females, UK citizens, all medically qualified who had undergone OSCE faculty training.

Tools

This research used video clips as 'tools' to help collect the required data. Two video clips of two medical students seeing a simulated patient were shown to every participant as a stimulus to further facilitate the process of data collection.

Before proceeding to the two videos, it is right to first make distinction in terms of describing the degree of acquaintance between perceiver/assessor and target/candidate. There are three levels of acquaintance (Kenny, 1994): zero acquaintance, short-term acquaintance, and long-term acquaintance. The first level of acquaintance is known as zero acquaintance where the perceiver and target do not meet, but the perceiver observes the target. For instance, a perceiver observes a target on TV. In zero acquaintance, perception is usually one-sided as the target does not see the perceiver. In this research, assessors were shown two video clips of students communicating with a simulated patient and were asked to observe and make judgements. Therefore, zero acquaintance was the level used in this study. Although zero acquaintance is not interpersonal, it can serve as an important baseline in the measurement of interpersonal perception (Kenny, 1994) and therefore can be used to answer my research question. The second level of acquaintance is known as short-term acquaintance where perceiver and target interact for a few minutes or even hours. This level represents what really happens in an ordinary OSCE station where a candidate is observed by an assessor for a few minutes and there is an actual interpersonal perception. The third level is known as long-term acquaintance in which perceiver and target know each other for a long time. This level might represent what

happens in the OSCE when assessors are teaching and examining the same students. The assessors in this level will have more information about the candidates and their skills and abilities.

The two videos (around 2 minutes each) featured scripted performances by real medical students from the University of Leeds (year 2 male student, and year 3 female student) in consultations with a middle-aged female simulated patient. The two videos were quite short because it might not be suitable and practicable to show two long videos (eight minutes each). In addition, each video has a start, a middle and a conclusion. In each video, the simulated patient depicted a new presentation to hospital. The two videos featured two different scenarios and cases: acute and unexplained loss of hearing in her right ear, and susceptibility and concern about diabetes. The two videos were recorded on two different days at the School of Medicine, University of Leeds, by professional technicians who work for the School of Medicine. The two medical students and simulated patient received formal invitations from the School of Medicine to take part in this project. The procedures and roles were described to them prior to the meeting. On the day of the recording, I met everyone separately to explain, in more detail, their roles and tasks and to answer any questions. They also had the chance to practice their roles on the recording day. They were asked to act normally as a medical doctor who sees a patient, but the female student was asked to show more concern and interest, while the male student was asked to show less attention (just to facilitate some discussions about communication skills). The two videos were saved on a portable tablet computer to be shown to the research participants in every interview.

Interview procedures

Before the commencement of my research interviews, I conducted four pilot interviews with four staff from the School of Medicine at the University of Leeds. These pilot interviews were necessary to build confidence, develop style, and gain experience and advice from experienced colleagues about data collection and 1:1 interviews.

Participants were first asked open-ended introductory and preliminary questions. Then, the first video was shown on the tablet computer and they were instructed to imagine that they were at an OSCE station. Participants watched the first video and were asked to write down and describe the student's characteristics and aspects of performance they considered essential to their judgement (I took field notes). Participants were presented with an OSCE mark rating sheet (Appendix 4) and asked to assess the student and justify their judgement by thinking aloud while forming it. They were also asked to describe the student in one or two words, and then they described the feedback they would give. Participants' thinking was then explored with follow-up questions before showing them the second video. The second video was then shown following the same previous procedures. After this, participants were asked a series of open-ended questions. Every interview was audio recorded and typically lasted 45-60 minutes.

Think aloud

Think-aloud protocol is a method that can help gather data in psychology and a range of social sciences where reading, writing and decision-making contribute to forming judgements. The participants in this research were required to think aloud every time they were assessing a student, and they were asked to say what they see, think and feel. Therefore, this method enables us to make explicit what is implicit (Ericsson & Simon,

1980) and therefore helps interpret and construct knowledge about how non-verbal behaviour influences assessors' judgements.

In this research, participants were required to deal with two different cases and deal with any difficulties they might confront in order to make an accurate decision. Some parts of this problem-solving process, analysing and making judgements are implicit because ''problem-solving means answering a question for which one does not directly have an answer available. This can be because the answer cannot be directly retrieved from memory but must be constructed from information that is available in memory or that can be obtained from the environment (for example, the givens of the problem or extra information that can be requested)... Therefore, problem-solving is the cognitive process to which the think aloud method is applied most frequently'' (van Someren et al., 1994, p. 8).

The aim of this research was to investigate and understand differences among assessors who observe the same performance. Olson et al. (1984) highlighted that using the think-aloud method could help investigate higher-level thinking processes and study individual differences in performing the same task. In addition, think-aloud data is considered a reliable source of information about thought processes (Ericsson & Simon, 1980).

Ericsson & Simon (1980) suggested three points as guidance to help maximise the efficacy of this method: (a) participants' active engagement, (b) participants describe their thoughts, and (c) shorten the time between participants' thoughts and their verbalisation. This guidance was followed in this research.

Triangulation

Attention to rigour has been emphasised in qualitative research, particularly on the state of medical education research (Britten, 2005; Wolf, 2004). In order to achieve rigour (Lingard and Kennedy, 2012) in this qualitative research, I emphasised (i) adequacy and appropriateness of the sample, (ii) the clarity of the analysis process, and (iii) the quality of the data collected. The latter ensures utilising techniques that could help capture naturalistic data. In order to collect valuable and validated data in my research, more than one method were utilised. Triangulation refers to the usage of more than one method in data collection as it facilitates data validation through verification from two or more methods (Bogdan & Biklen, 2006). When two or more methods provide the same result, there is more confidence in the result.

In this research, data was collected using more than one method. Firstly, participants were interviewed individually with interviews using a familiar and well utilised method of data collection in qualitative research (Dicicco-Bloom & Crabtree, 2006; Harris, 2002). This method was used because it helps obtain participants' personal perspectives and experiences on different topics (Crabtree & Miller, 1999). I used what is known as 'depth interview' because it is identified to provide rich and detailed relevant information (ibid). A semi-structured format was used in the interviews. This format was guided by predetermined open-ended questions with freedom to pursue additional related topics when needed (Dicicco-Bloom & Crabtree, 2006). Secondly, the participants observed performances and made judgements using think-aloud protocol. Thirdly, the participants were asked to write down and describe students' characteristics and aspects of performance. Ericsson and Simon (1980) highlighted that think-aloud data from working memory cannot always be complete and that a number of thought processes would be

excluded from being expressed verbally. In addition, person perception, as mentioned earlier, relies to a great extent on associating people with traits. A large number of studies in person perception ask perceivers to rate the targets on scales and rarely are perceivers asked to provide a free description of the target (Kenny, 1994). Therefore, every assessor in this research was asked, after watching the video clips, to write a free description about the students and their performance. This could help enable the assessors to provide more details and elaboration on observed behaviours and traits that might not be mentioned in some scales. They could also write down some notes and then describe them in detail verbally.

Analysis and coding

Data analysis in my research helped make sense of what had been gathered. Audio recordings were transcribed word for word and checked for accurateness. The collected data was not analysed in one setting. Rather, it was an ongoing process of transcribing, reading and reasoning the meaning of the data as they were being gathered.

Following repeated reading, codes were assigned (Appendix 5). Coding is simply described as ‘‘the process of categorizing and sorting data’’ (Charmaz, 1983, p. 112) in order to identify instances that are similar in concept - thematic analysis (Strauss & Corbin, 1998). Reading the transcripts required me to demarcate segments within it, and each segment was labelled with a code (Saldana, 2013). Transcripts were first conceptualised line-by-line, known as open coding (Strauss, 1987) to allow ‘‘the process of breaking down, examining, comparing, conceptualizing and categorising data’’ (Strauss & Corbin, 1990, p. 61) to take place. This process enabled gradual building up of major categories I found. Codes were first grouped as ‘‘assessor related codes’’, ‘‘student

related codes'', ''patient related codes'', and ''other ambient related codes''. Later, axial coding was employed which involved putting data back together in a new way by making connections between major categories and subcategories. Strauss and Corbin (1998, p. 123) defined axial coding as ''the process of relating categories to their subcategories, termed 'axial' because coding occurs around the axis of a category, linking categories at the level of properties and dimensions''. This could help put ''the fractured data back together in new ways after open coding, by making connections between a category and its subcategories'' (Strauss & Corbin, 1990, p. 96).

Notes or memos were taken whenever needed during the process of data analysis to highlight the relationships between codes and ideas. Glaser (1978, p. 83) defined memos as ''the theorizing write-up of ideas about codes and their relationships as they strike the analyst while coding''. Highly structured data, from tightly defined questions such as participants' one word judgements and impressions, was coded as a layer on top of the data without added segmenting of the content.

Ethics

Ethical approval (Appendix 6) was obtained from the University of Leeds before the commencement of data collection. Every participant was sent an information sheet (Appendix 7) that described the aims of the project and the procedures of the interview. The information sheet was sent with the e-mail invitation to all participants. Every participant signed a consent form (Appendix 8) that confirmed anonymity and confidentiality. In addition, the two medical students and simulated patient were aware of their roles and they also signed different consent forms (Appendix 9,10).

Chapter 4 Results

The previous chapter described the methodology and research philosophy chosen and applied to conduct this research. This chapter details the findings about what and how non-verbal behaviour influences assessors' decisions when assessing medical students using the OSCE as an assessment instrument.

In reaching theoretical saturation, 18 participants were recruited to take part in this study, 11 males and 7 females, UK citizens, all medically qualified who took part in assessing medical students at the University of Leeds using the OSCE. They all had attended a training course about how to assess candidates' competence using the OSCE.

The assessors in this research had different experience and impressions about the OSCE as an assessment tool. Table 5 shows the experience of each assessor along with their impressions of the OSCE.

Table 7 Assessors' experience and impressions

Assessor	Experience	Impression of the OSCE
Assessor no. 1	2 years	''It is busy. I would not say I enjoyed but satisfied. Not the best format but the best achievable. I have been in both sides of the desk there.''
Assessor no. 2	5 years	''it is good and interesting.. Not ideal.. I was a student myself, I know it is stressful''
Assessor no. 3	12 years	''I enjoyed the experience seeing different levels''
Assessor no. 4	4 years	''I experienced it as a student.. The content of the exam was very predictable.. It does not stretch the students as much as it should''
Assessor no. 5	7 years	''It is fun really.. Bring me a little bit up to date, because that is not what I do on the daily basis''
Assessor no. 6	14 years	''The most frustrating thing about the OSCE is trying to not say anything above and beyond what is in the script and trying to maintain a consistency with each candidate that comes through''
Assessor no. 7	3 years	''I had an OSCE as a student as well.. I remember it was terrifying but I thought it was a very fair way to do it.. It is stressful for the students, I know that; I experienced it''
Assessor no. 8	2 years	''Generally it has been a fairly good experience.. I was nervous when I first started assessing them''
Assessor no. 9	6 years	''I find it very enjoyable.. Sometimes the tasks that are asked of the students are a little unrealistic''
Assessor no. 10	5 years	''I find it interesting.. I have learned things from the OSCE.. I usually find it pretty straightforward''
Assessor no. 11	7 years	''I enjoy it. I find it interesting. The students who come through are very different, and I like to hear how much the students have learned. A lot of them I have met before, so it is nice how they are progressing through the years''
Assessor no. 12	5 years	''It is not like real, but it is effective''
Assessor no. 13	5 years	''I enjoy it''
Assessor no. 14	1 year	''An OSCE itself is a broad assessment method. Different OSCEs are suited to different things''
Assessor no. 15	2 years	''It is relatively stressful at first, the beginning of each station, but generally quite enjoyable''
Assessor no. 16	6-7 years	''Good way of differentiating students.. The challenge is trying to integrate different skills and do that in a way that can be examined in 8 minutes''
Assessor no. 17	4 years	''I find it interesting to do it.. Sometimes you learn things''
Assessor no. 18	11 years	''Interesting and realistic''

As detailed in the previous chapter, every assessor was shown two video clips of two medical students seeing a simulated patient. The assessors were asked to assess the two students using grade descriptors (Appendix 4). The results of assessing the two medical students clearly showed a quite wide variance in making judgements. The following table

shows each assessor's global judgement to the performance of the two medical students in addition to their impressions about each student.

Table 8 Assessors' global judgements and impressions of candidates' performances

Assessor	Gender	Global judgment Student 1	Impression	Global judgment Student 2	Impression
Assessor no. 1	Male	Clear fail	Disinterested**	V. good pass	Excellent
Assessor no. 2	Male	Clear fail	Appalling	Borderline	Receptive
Assessor no. 3	Male	Borderline *	Casual	Borderline	Ineffectual
Assessor no. 4	Male	Clear fail	Rude	Clear pass	Kind
Assessor no. 5	Female	Borderline *	Disinterested**	Clear pass	Sympathetic
Assessor no. 6	Male	Borderline	Disinterested**	Clear pass	Professioned***
Assessor no. 7	Female	Clear fail	Uncaring	Borderline	Pleasant
Assessor no. 8	Female	Borderline *	Arrogant	V. good pass	Empathetic
Assessor no. 9	Male	Clear fail *	Disinterested**	V. good pass	Empathetic
Assessor no. 10	Female	Borderline *	Uninterested**	V. good pass *	Engaged
Assessor no. 11	Female	Clear fail *	Poor.communic-	Clear pass *	Personable
Assessor no. 12	Male	Clear fail	Unprofessional	Clear pass *	Supportive
Assessor no. 13	Male	Clear fail *	Unpleasant	Excellent *	Professional
Assessor no. 14	Male	Clear fail *	Poor	Clear pass *	Fine
Assessor no. 15	Female	Borderline	Unprofessional	Clear pass	Competent
Assessor no. 16	Male	Clear fail	Poor	Clear pass	Average
Assessor no. 17	Female	Clear fail *	Unprofessional	V. good pass	Open
Assessor no. 18	Male	Borderline *	Unaware	Clear pass	Smiley

* The assessors gave two decisions (e.g. between borderline and pass) before they decided to go with only one. ** The student showed 'a lack of interest'. *** ''By professioned I mean came and did the job, but there really was not much extra to it, but was at the line you might expect for somebody at their level.''

Such differences in judgements might not be very surprising to some assessors as they made it clear that there will always be a subjective element in the OSCE. This subjectivity varies from one assessor to another.

Yes, you could achieve the task, you could put all the shapes in the shape sorter or whatever and that is it but I think there is a subjective element. Because medicine is a social skill amongst other things I think there is bound to be a subjective element. There is a subjective element in the overall global impression'' (assessor no. 9)

''It is quite objective, but there are subjective elements as well. There is a bit both because a lot of it is how the students come across and their communication. This is not just about, I suppose, ticking the boxes'' (assessor no.11)

This study focused on increasing inter-rater reliability by understanding and then decreasing such variance and subjectivity. The focus was on non-verbal behaviours that could influence the assessor's judgement.

Thematic analysis revealed a rich framework where the interaction and non-verbal behaviours of assessors, patients and candidates all contributed to global ratings. Assessors' identification and response to candidate behaviours was complex and individual. Subthemes included several elements that could highlight different non-verbal behaviour of each character in the station (student, assessor and patient). All three characters could have been influenced by elements relating to how the exam was organized and prepared (Figure 6,7).

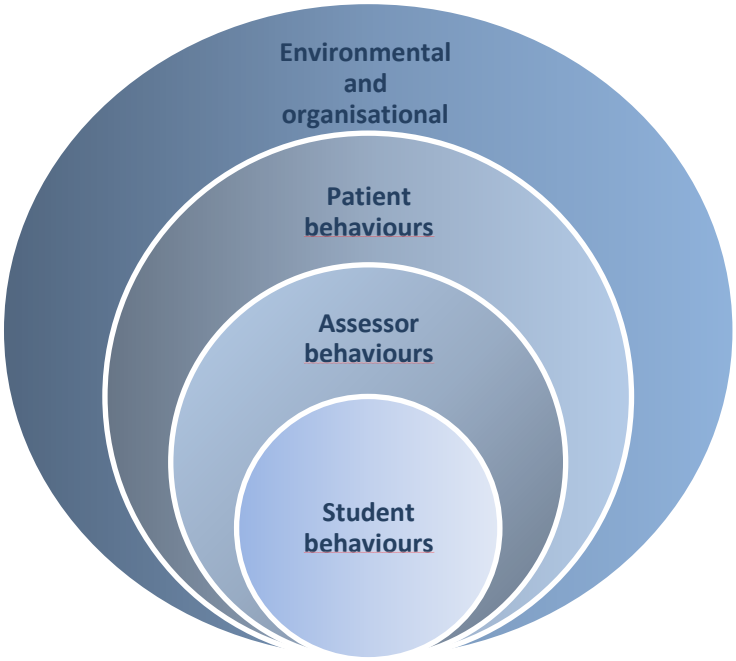

Figure 6 Main themes

Figure 7 Subthemes (each element will be explained later)

A- Student related behaviours

A-1- Bedside manner

The first non-verbal behaviour that has the potential to influence assessors' global marking is the student general manner or what is known as 'bedside manner'. Some assessors used the word 'manner' to give a general perspective of how a student might behave while seeing a patient. For example:

''seemed casual in manner.. They do not need to be an excellent student in order to pass.. If somebody has a good manner and a basic sound knowledge, that is fine'' (assessor no. 3)

However, it was ideal to make it more specific and understand what this term (manner) might refer to. Six major points, see Figure 8, were identified that the assessors had used to refer to bedside manner.

Figure 8 Bedside manner

A-1-1 Listening

The interviewed assessors in this research made it very clear regarding the necessity of listening in order for a student to graduate and become a doctor. Some assessors directly commented, after watching the two videos, on the presence of such an essential skill and encouraged it:

''she had good listening skills''(assessor no. 1).

''she did listen to the patient'' (assessor no. 5).

Other assessors noticed the absence of the required listening skills and blamed the student for not being able to listen to what the patient was trying to say. For instance, they stated in their feedback that:

''He needs to be paying attention to the patient, and appear to be listening carefully, and to give her more opportunity to ask things or mention things''(assessor no. 12)

Some assessors referred to listening indirectly by encouraging the students to let the simulated patient talk and explain whatever she wanted to say, and to give her enough time to ask questions and describe concerns:

''She allowed the patient to talk''(assessor no. 6)

''How much time they give patients to answer their questions'' (assessor no. 17)

Other assessors made a general statement about the necessity of listening, and made it an important feature that doctors need to possess:

''When I was a patient I wanted my doctor to be more receptive'' (assessor no. 2)

''They (students) need to be attentive and listen to what the patient wants to say'' (assessor no. 1)

''I think it is important to have good listening skills'' (assessor number 8)

Finally, some assessors highlighted the influence of some factors, anxiety for example, on listening. Such factors can have a negative impact on their listening skills.

''The difficulty for them (students) is the listening part, particularly the second years where they are trying to think what to ask next, and I think that can affect their listening skills'' (assessor no. 8)

A-1-2 Showing respect and interest

The second point on how the assessors described 'manner' is related to how the student respected the patient and showed interest. Some assessors described the lack of respect and pointed it out whenever there was the opposite of respect, e.g. rudeness, carelessness or arrogance:

''He was rude.. he did not show respect'' (assessor no. 2)

''He was uncaring and disinterested.. loafing.'' (assessor no. 7)

''He was arrogant, unbothered, disinterested'' (assessor no. 8)

''There was a background of rudeness and impoliteness actually''(assessor no. 12)

Some assessors described some actions and behaviours that students did as off-putting. Such behaviours are not considered to be polite and hence the patient might not feel respected. For instance, playing with a pen or chewing were not acceptable.

''He was playing with his pen'' (assessor no. 1)

''Chewing and fiddling'' (assessor no. 11)

''Respect the patient rather than looking distracted playing with his pen'' (assessor no.15)

''While taking notes they asked questions which is generally seen as rude behaviour by some'' (assessor no. 18)

Furthermore, the assessors mentioned that one way to respect a patient is by choosing your words carefully. This refers to how you generally communicate and choose your style to ask questions:

''His communication skills were dreadful'' (assessor no. 7)

''Some pieces of the advice such as 'use the phone in the other ear you can hear with' could be interpreted as flippant'' (assessor no. 18)

Some assessors connected respect with how the student managed to establish rapport with the simulated patient and made her feel. For instance, some assessors emphasised being warm and polite.

''She is polite and warmer than the first student'' (assessor no. 3)

''Establish rapport and treat patients with respect and dignity..'' (assessor no. 9)

''He did not interact properly.. a little bit patronizing'' (assessor no. 9)

Finally, respecting the patient was seen as very essential and important regardless of how the student felt about the patient. It is something the student must do:

''He was disinterested! He needs to be interested in what he is doing! Or at least to show that he is interested'' (assessor no. 1)

A-1-3 Sympathy

The third point that was among the bedside manner features was sympathy. The assessors wanted the students to be sympathetic. The patients come to the hospital or clinic to feel supported and to share their feelings with their doctors. A relationship of harmony and affinity is what the patients expect to experience when they consult those who are given

the responsibility to take care of them. Some assessors in this research directly encouraged the presence of sympathy:

''She was sympathetic'' (assessor no. 5)

''Lack of empathy'' (assessor no. 5)

''it is important to show empathy as well'' (assessor no. 8)

Other assessors wanted to observe concerns elicitation as a way of showing sympathy which would allow the patient to feel encouraged to say what bothers them and feel supported:

''They need to elicit the patient's concerns'' (assessor no. 1)

''Try to give the feeling she is actually being listened to'' (assessor no. 11)

''She was supportive'' (assessor no. 12)

''She was very reassuring.. and sympathetic'' (assessor no. 13)

A-1-4 Body language

Body language is one way to communicate non-verbally. Thoughts and feelings of candidates are possibly expressed by such a type of communication. The assessors highlighted the importance of assessing body language because it is a part of communication.

''The video makes it difficult to assess the body language'' (assessor no. 2)

''good communicators usually have body language and facial expressions'' (assessor no. 11)

Body language was referred to, by the assessors, as several behaviours expressed physically such as body posture, eye contact, gestures, body space, and facial expressions.

''Body posture and eye contact are more important'' (assessor no. 4)

''Being close to the patient without invading their personal space'' (assessor no.9)

''He did not look at the patient.'' (assessor no. 11)

''She was leaning forward and looking at the patient.. and smiling'' (assessor no. 11)

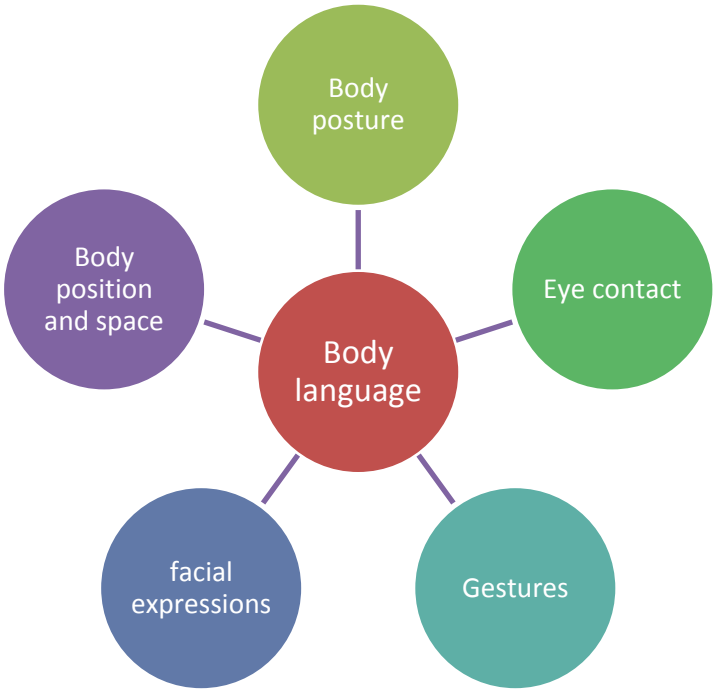

Figure 9 Body language

A-1-5 Dedication

It was clearly identified that all doctors need to dedicate their time to their patients. For instance, some assessors marked down the students because they looked at their watches while seeing the simulated patient:

''The thing it marked it down and it did not make for me to make it very good pass was the slight lack of professionalism when she was looking at her watch, that would be off-putting to the patient.'' (assessor no. 5)

When a patient comes to see a doctor, he or she expects to receive a professional service because it is related to their health. This requires complete focus on the patient. Otherwise, the patient will not feel comfortable.

''The patient sometimes wait for a long time to be seen by a doctor… One of my relatives was annoyed when her doctor was in a hurry'' (assessor no. 2)

''She was very overtly looking at her watch which is fine, but if you explain to the patient why you doing it.. But when you keep doing that it looks like you are ready for your dinner and that is a little bit off putting'' (assessor no. 9)

A-1-6 Establishing rapport

It was highlighted that the relationship between a patient and his or her doctor needs to be close. Any consultation should be built on trust and understanding. Rapport is usually established at the beginning of any consultation by introducing each other and creating healthy and friendly atmosphere.

''She established rapport and she was interested to hear what the patient wanted to say'' (assessor no. 3)

''She was approachable ... She introduced herself well.. There was good rapport'' (assessor no. 7)

A-2 Adaptation

It was highlighted that it is important for medical students to demonstrate the ability to adapt from one situation to another. This adaptation ability is essential in their future career when they see different patients with different characters, needs and concerns. It would be hard for any medical doctor to collect information and data if they cannot adapt themselves to different situations properly.

''They need to get information in a way that is appropriate to the patient. The way can be different from one patient to another.''(assessor no. 1)

''I look if the students can adapt themselves to different patients because different patients have different worries.. different styles.. Someone's precise information and details.. and other lots of information requires such a skill'' (assessor no. 1)

A-3 Patient involvement

The assessors in this research made it very clear about the importance of treating the patient as a human being and equal partner. This requires adequate engagement with the patient.

''He did not really engage with the patient in any way'' (assessor no. 4)

''One thing that I look for in candidates is that I want to see them treating the person as well as the disease kinda thing'' (assessor no. 4)

''Treat the patient as an equal partner in the negotiation'' (assessor no. 9)

In order to achieve such adequate and appropriate engagement and involvement, the students need to be aware that they are not intended to give the patient orders or dominate the consultation.

''If candidates whose approach is like 'me doctor you dog'.. so do not ring the right bells for me'' (assessor no. 9)

''Do not dominate the consultation'' (assessor no. 12)

Rather, the student is encouraged to discuss and reach a mutual agreement with the patient. The patient needs to understand and agree on what should be done. Such involvement helps maximise the benefits and efficacy of the consultation.

''Does not seek to come to a mutually agreed plan of action'' (assessor no. 16)

''The plan did not involve the patient'' (assessor no. 18)

''Did the student come up with a reasonable explanation of what the problem was? Did you share that with the patient? Did you gain the patient's agreement from what is going on? Did you come to a satisfactory conclusion whereby you both agree to take this forward'' (assessor no. 18)

A-4 Emotional status

The emotional status of the student during the station played a role and influenced the assessor's global marking.

''In video you do not get any kind of emotional connection with the student'' Was he nervous, happy or excited? ''Which I think is an important part of consultation skills in medical practice'' (assessor no. 4)

First, such emotional status can affect the performance of the student him/herself, which will in turn affect the assessor's judgement.

''There is less chance for the students to dig themselves out of the hole if they had a disastrous station earlier on that day'' (assessor no. 1)

''One station can affect the student in the next station by being pressured and thinking about performance if went well or not'' (assessor no. 1)

Second, some assessors had experienced the OSCE when they were students. This experience helped them to better understand the emotional status of the students during the examination. Therefore, whenever they see a nervous or stressed student they would have a better understanding of the causes of such emotional status.

''When I was a student, I remember it was terrifying but I thought it was a very fair way to do it.. It is stressful for the students, I know that. I experienced it'' (assessor no. 7)

Third, some assessors might tend to sympathise with the student when he or she is stressed. Such sympathy may not necessarily directly reflect on the assessor's judgement. Rather, it could be reflected in the leeway the assessor gives the student which in turn could result in better performance, and then better achievement and a higher score.

''When I say stressful, it is more that you want to encourage the student, to support the student to give all the facts and the knowledge that they have, but sometimes it stressful because you are not allowed to give too many prompts or too much information. So, if it is a

nervous student and you are trying to push them on the right direction, that can be stressful'' (assessor no. 15)

A-5 Student's knowledge and skills

Unsurprisingly, and since OSCEs are about showing competence and skills, the student needs to show competence, knowledge and skills when they are being observed during an OSCE.

''They do not need to be an excellent student in order to pass.. If somebody has a good manner and a basic sound knowledge, that is fine.'' (assessor no. 3)

''Good communicators usually have confidence, knowledge, body language and facial expressions'' (assessor no. 11)

''When you are assessing students, it is good to make sure that some ideas of basics there because you need the basics and the structure and the coverage of the relevant issues'' (assessor no. 12)

The assessors were keen to see the amount of knowledge a student can manifest and apply. They were impressed when there was good basic knowledge, and it was annoying to them when the student did not adequately apply knowledge at the station.

''Seems to know what she was asking and what she was talking about'' (assessor no. 13)

''It is more that you want to encourage the student, to support the student to give all the facts and the knowledge that they have'' (assessor no. 15)

''He did not actually have a chance to apply knowledge because he did not give the patient any chance'' (assessor no. 18)

The assessors highlighted that the absence of enough knowledge would reflect on the data collected and history taking which could in turn affect the diagnosis and treatment plan.

''He did not even look in the ear'' (assessor no. 7)

''The information gathering should have included an exam'' (assessor no. 18)

A-6 Confidence

Confidence was seen by some assessors as a requirement that students need to manifest when seeing a patient. They were keen to observe candidates confidently talking to patients and communicating with them.

''The candidate has to confidently identify who they are, and what they there for, both to me and to the patient'' (assessor no. 9)

''Good communicators usually have confidence, knowledge, body language and facial expressions'' (assessor no. 11)

''The green card.. to a very confident student'' (assessor no. 17)

It was identified that confidence helps candidates perform and achieve better because they are relaxed and more focused. In addition, some assessors highlighted that confidence usually comes with good and high qualifications and vice versa.

''Older students will pretend to be understanding a question if they really do not.. because they do not want to show that they are not qualified.. so they make up things that are entirely wrong. On the other hand, younger students will find it easy to say I did not understand'' (assessor no. 3)

''She was a bit wishy-washy'' (assessor no. 7)

''If they are calm they can do it well, but if they get nervous they… It is quite difficult to spot the difference between somebody who has got a clue and somebody who has not once they get nervous'' (assessor no. 7)

Finally, it was mentioned that when the student and the examiner know each other, this could decrease the level of stress that the exam usually places on students, therefore they could perform better.

''They might feel more relaxed with somebody they know as an examiner'' (assessor no. 11)

There was no evidence found of the reverse (i.e. knowing the assessor made it more difficult or stressful). Such evidence would probably require a direct question to candidates. However, assessors themselves might find it difficult to assess a candidate they know, as will be described later.

A-7 Appearance

The way the student dressed was highlighted in detail by some assessors. They mentioned that, generally, it would be expected that a medical practitioner wears professional dress and uniforms. Otherwise, they might cause some distraction.

''Inappropriate dress can be distracting'' (assessor no. 8)

''That was unprofessional dress'' (assessor no. 1)

''She is professionally dressed'' (assessor no. 1)

Some assessors went into some details about what clothes would or would not be acceptable from a medical and professional point of view.

''I do not expect to see a student wearing a colourful shirt and a jeans'' (assessor no. 2)

''Very smartly dressed, he looked the part'' (assessor no. 11)

''The dress sense was probably inappropriate in that t-shirt is not generally acceptable'' (assessor no. 18)

The smell was mentioned as another point that is related to the appearance of the student during an examination. Both dress and smell are expected to be professional and acceptable.

''Nonverbal communication such as clothing, smells.. catch my attention'' (assessor no. 1)

Finally, attractiveness was mentioned, by one male assessor, and it was highlighted that it could affect the judgement of some assessors. However, it was clearly stated by him that such a thing should never be considered as a criteria.

''I remind myself of the tendency for medical examiners to give attractive women better marks.. I need to focus on their skills.. to be fair'' (assessor no. 1)

A-8 Random vs ordered performance

The assessors emphasised the importance of following a logical order and approach when seeing a patient. They criticised the students whenever they felt that there was some randomness in their approach and style.

''He was disorganized'' (assessor no. 1)

''The questions were random'' (assessor no. 4)

''He was not organized.. He was random from one idea to another'' (assessor no. 7)

''Random approach jumping backwards and forwards'' (assessor no. 16)

It was highlighted that such randomness in style and approach needs to be seen as a deficiency and therefore reflected in the student's mark.

''a candidate who simply just spits everything out in a disordered fashion but still manages to get marks on an OSCE station should not score as highly as a candidate who goes through the question in an orderly manner'' (assessor no. 6)

It is important though to highlight that it was clarified that there is a difference between randomness and flexibility. As mentioned earlier, the student is expected to show some level of adaptation. This adaptation requires flexibility and the ability to manage the station differently from one patient to another. However, this adaptation and flexibility can still be predictable and planned.

''She was not very structured.. but she was flexible'' (assessor no. 10)

A-9 Concentration and distractors

Student concentration and paying attention during the station was considered another important factor the assessors thought necessary and wanted the students to have. Whenever concentration was influenced, the assessors made a comment.

''He was easily distracted'' (assessor no. 1)

''He was not focused'' (assessor no. 2)

''He did not pay attention to what the patient said'' (assessor no. 18)

Not paying attention is seen as annoying not only by the assessors, but by the patients as well. Both expect the student to pay attention and have concentration as a part of adequate engagement.

''Patients don't like you not paying attention to them'' (assessor no. 18)

Any issue with concentration during the station was possibly shown to be reflected on the mark awarded to the student.

''The reason I do not think it is a very good pass is because she was still occasionally distracted'' (assessor no. 6)

Concentration might be influenced by several distractors. The assessors in this research described what the students in the two videos did that could influence and distract their attention and concentration. They also mentioned similar examples that could influence concentration. Therefore, such distractors were discouraged. It is important though to note that the following distractors are just examples selected by the assessors in this study.

''He was playing with his pen'' (assessor no. 1)

''She looked at her watch twice.. She looked at her phone once'' (assessor no. 6)

''Fidgeting and playing with hair''(assessor no. 8)

''Fiddling with his pen, writing notes and waving at someone else'' (assessor no. 9)

''Writing a lot down'' (assessor no. 16)

It was suggested that if a student had to look at his or her mobile or watch, they need do it more subtly and tell the patient the reason so the patient can understand why the student is doing it.

''Patients don't mind you looking at your watch, looking at the computer, looking at a book, what they don't like is they don't like you not paying attention to them.. If you are going to

check the time or your phone, you need to do it more subtly.. You need to explain to the patient why you are looking at your watch'' (assessor no. 18)

Figure 10 Distractors

A-10 Struggling with role play

As mentioned earlier in Chapter 1, the OSCE is a type of examination that is conducted in an artificial way. There are simulated patients or actors who play different roles.

''There is no secret an OSCE is removed from reality to some extent'' (assessor no. 4)

''The problem with simulated patients is that you know it is a simulated environment, and you can never get away from that'' (assessor no. 14)

It was highlighted that some students may struggle with the role play concept when they are asked in an artificial situation to treat simulated patients who are asked to give specific and not genuine responses and reactions.

''There is a group of medical student who struggle with the role play concept, that they find it difficult to take on the role that you are asking them in an artificial situation'' (assessor no. 6)

''I think the simulated patients are quite often realistic, but sometimes you cannot get an actual spot on simulated patient. The students sometimes have trouble suspending disbelief'' (assessor no. 9)

''I think the students are aware of when they doing it with the simulated patients, and they are aware that they may.. will be, if you like, elements of the simulated patient is told not to tell them, whereas the real patients I think they are just informal to actually just give them genuine response each time'' (assessor no. 18)

This difficulty with the role play concept can cause some levels of anxiety to the students which in turn could affect their general performance.

''Those individuals who struggle with role play tend to be more nervous, more anxious'' (assessor no. 6)

Not only does this struggle with the role play concept influence the student performance, but the assessor in this situation might find it difficult to handle, especially when it comes to assessing certain abilities and skills.

''Those individuals who struggle with role play tend to be more nervous, more anxious.. it is difficult to assess their knowledge.. It is even more difficult to assess their transferable skills'' (assessor no. 6)

This struggle though is mostly going to decrease with time as the students will progress and get more familiar with the process and experience. Therefore, the influence of the role play concept and its influence might be more noticeable in the early years of study students.

''In the first experience of OSCEs, the students tend to panic and they do not read the information that they are given before they come into the station because they are too nervous, and usually as years go on that changes, they are less nervous. And they have got into the roleplay of what they need to do because there are certain things that just very much roleplay such as introduction to patients, hand washing..'' (assessor no. 15)

A-11 Reasoning, synthesis and planning

One point the assessors looked at and wanted the students to possess and show is how they can reason, synthesise information and plan for next steps and procedures. Reaching a diagnosis and treating a patient requires such skills. Although these skills can be implicit, they would be manifested and assessed in how the students take the patients through the station and answer their questions.

''She showed good synthesis of information.. and she had a plan'' (assessor no. 1)

''He was uncertain.. He did not show any evidence of reasoning'' (assessor no .8)

''She was listening and synthesising what the patient said, repeating back what the patient said.. managed the questions very well'' (assessor no. 17)

A-12 Questioning and thoroughness,

The assessors highlighted the importance of data collection and information gathering. Although it is more verbal, such a skill requires good listening, communication and questioning and answering abilities. In addition, it requires thoroughness to cover everything required to be covered to reach an accurate diagnosis and treatment plan.

''I like them to be thorough'' (assessor no. 4)

''The green card is about being exceptionally thorough'' (assessor no. 12)

Part of thoroughness includes understanding and exploring the patient's concerns and worries before proceeding directly to more detailed questions.

''My criticism would be that she launched directly into a long answer of the patient's question rather than spending time exploring the patient's underlining concerns'' (assessor no. 4)

''She did not explore why the patient was there.. She failed to explore ideas and concerns.. and did not ask open questions'' (assessor no. 7)

It was highlighted that in order to be thorough the questions should be open so the patient can answer them freely.

''questions were not open, they were directive.. he did not ask broad questions whereby the patient would actually give him information'' (assessor no. 18)

''She did too much on the social waffly bit about keeping a happy life style and silly comments like that. It was softer not hard questioning'' (assessor no. 15)

In addition, when a student provides information to the patient, it needs to be thorough. The given information should cover everything needed to be covered.

''She did not go specific in terms of how much exercise, how often, for what length'' (assessor no. 16)

Probing was another important point that a student needs to do in a station. It is not about a list of questions a student asks a patient. Rather, it is about looking for details and clarification.

''Whether they follow the information they get from the patient with the appropriate next questions'' (assessor no. 15)

When asking a question, choosing the words carefully plays a role in giving the patient a chance to answer them optimally. It is about a skill of choosing the right words to make a question.

''She could alter how she asked that final question . She could say rather than 'any more questions' to which people tend to say 'no', you could say 'what questions do you have for me now?' 'What else do you like to know' (assessor no.18)

Finally, concluding questions are important to let the patient say anything they might have forgotten to ask. It also increases thoroughness as new information can be obtained.

''She asked some further questions at the end to see whether there is anything the patient could ask'' (assessor no. 18)

A-13 Fluency

It was identified that fluency, both of speech and performance, was an important point that could influence assessors' judgements. It is impressive when both performance and speech go smoothly without many interruptions and hesitations.

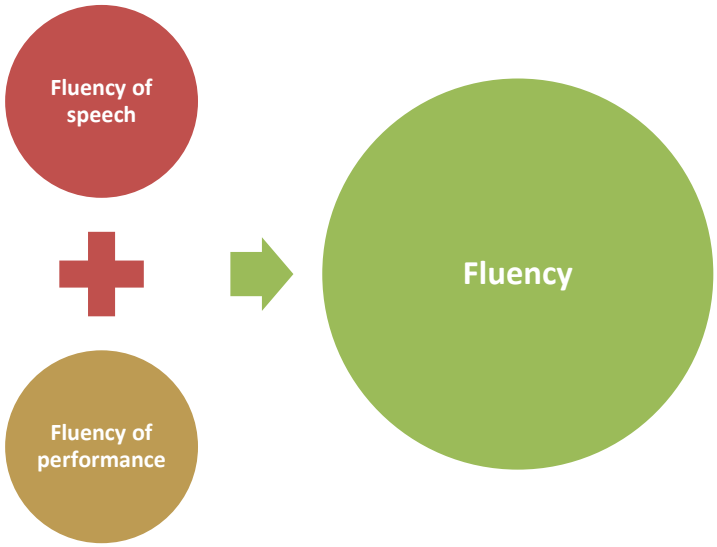

Figure 11 Fluency

''I suppose the most challenging thing is if I find them (students) hard to follow their flow of speech, and that would distract me'' (assessor no. 10)

''So, that makes it slightly a sort of stilted.. a slower consultation.. Jumping backwards and forwards, he does not flow nicely'' (assessor no. 16)

''If it flows seamlessly it impresses me much more than if you can hear the bells and whistles going round in the candidate's head'' (assessor no. 9)

A-14 Culture-related behaviours

It has been found that the culture of the student may play a role in how they perform in an OSCE station, and in turn influences the assessors' global marking.

''I think that different cultures do present a challenge sometimes'' (assessor no. 4)

First, students from different cultures who speak English as a second language may struggle with some linguistic issues and the way they communicate with native speakers. This issue was highlighted by some of the assessors in this research.

''I think there are issues with people who have English as not their first language, or people who have English as their first language but are part of another culture where they have a second language at home, they are not English culture (assessor no. 4)

''People from different cultures may sometimes have linguistic barriers in assessing people from a different culture in a language that is not their own language'' (assessor no. 12)

One thing related to this linguistic issue is the ability to understand what the patient is saying. Non-native speakers might find it difficult to understand every single word said by the patient. This could in turn affect their understanding of the case and their further steps.

''If they speak English fluently then in general their communication skills will be better. If their English is something that they are struggling with then in my observation the interaction would be more difficult… Sometimes the patient may struggle to understand what they are saying, or may struggle to understand the response'' (assessor no. 10)

''I think sometimes they can miss the subtleties of what the patient is saying because there might be differences in phrases or mannerisms'' (assessor no. 16)

This language understanding issue might be less noticeable with simulated patients compared to real patients because simulated patients are more trained to deal with such cases and they can paraphrase their responses.

''If English.. it clearly is not their first language, then sometimes patients or the simulated patients can say things and they do not always necessarily pick up or understand the phrase that can be used. But usually simulated patients are good enough so they can rephrase it'' (assessor no. 5)

''If it is a real patient there might be language barriers because they might use terms a foreign student would not be aware of'' (assessor no. 15)

In addition, the linguistic issue could affect the students in how they send a message and talk to the patient. This possible struggle can affect how they probe and ask further questions in order to reach an accurate diagnosis and treatment plan.

'''I think they can struggle in the reverse (sending instead of receiving) in how they give information because again of that they happen to translate in their mind'' (assessor no. 16)

Furthermore, some accents can be heavy and difficult to follow and understand.

''Some international students although their language is grammatically correct, they have heavy accent which may not be easy for the assessor or the patient to understand'' (assessor no. 18)

Second, some students might struggle, as mentioned earlier, with the role play concept. This struggle can cause the student to be more nervous and anxious during the examination. It was mentioned that students from different cultures struggle more than local students with the role play concept.

''The background and experience of the candidate does have a bearing with their reliability to do role play. My experience has been that there are individuals who are struggling to get to grips with the role play roles, boundaries, and how it works. It is much more likely that those candidates are probably from overseas or foreign background'' (assessor no. 6)

''Those individuals who struggle with role play tend to be more nervous, more anxious, and come from a background which is predominantly overseas'' (assessor no. 6)

Third, it has been found that there are some culture-related behaviours that might cause an issue during the station. A misunderstanding can happen when a student does not do, because of some cultural differences, what the patient and assessor expect him or her to do. The majority of the assessors in this research were aware of such cultural differences.

''Assessors are looking at attitude.. Perhaps female students from some countries may find it inappropriate to keep eye contact with a young man'' (assessor no. 18)

''I can understand the difficulties of minority ethnic girls about shyness'' (assessor no. 9)

''There can be cultural differences that make a difference in the consultation.. it is possible that people from some backgrounds have ideas of patients as less being individual, less worthy of respect as individuals. I have seen that happen'' (assessor no. 12)

However, sometimes it would not be clear to the examiner whether a certain behaviour was related to the culture of the student or not. The examiner is not allowed to ask the

student extra questions during the station. This is why different examiners might deal with such scenarios differently.

''I would not give the student full mark, for example, if she refused to shake hands with a male patient'' (assessor no. 12)

''You would not know because we are not asking them.. Was it just you are rude or not interested, or actually there is a cultural difference.. Generally, the examiner would not interact with the student'' (assessor no. 16)

Fourth, different cultures can have different meanings to different behaviours and even voice tones which could be sometimes missed or interpreted and perceived differently by both the patient and the student.

''Some of the body language may be different in different cultures and that can be perceived differently by the patient'' (assessor no. 16)

''The way some patients behave and the way they want response to them is hard to pick up if you are not used to picking them culturally'' (assessor no. 14)

''It is often more difficult for them to.. I suppose because sometimes the tone of somebody's voice has an effect on a patient'' (assessor no. 11)

Fifth, and as mentioned earlier, the fluency of speech might influence the assessor's judgement. Students from different cultures who speak English as a second language can have an issue with speech fluency because they keep translating what they hear and what they want to say from their first language to English. This may cause some slowness in speaking and responding.

''Certainly students who have come from overseas to train can struggle sometimes with the consultation skills. Some of the things they struggle with is language. So, they happen to sometimes translate what the patient is saying back into their own language. So, that makes it slightly a sort of jilted.. a slower consultation'' (assessor no. 16)

In addition, some female and male students might feel pressured and embarrassed talking about sensitive topics with the opposite sex, and this might vary from one culture to another.

''Females might find difficulties assessing men on sensitive topics'' (assessor no. 9)

''Perhaps male students struggle a little bit more with female patients if they are talking about something embarrassing or something that gynaecological related or contraception.. if there is a male patient coming about some male problem then the females struggle a little bit with the consultation'' (assessor no. 11)

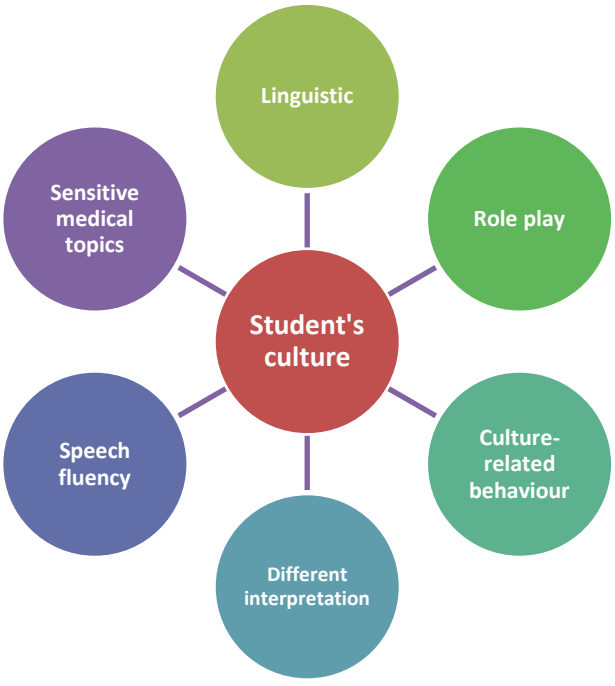

Figure 12 Student's culture

A-15 Safety assurance

Since the students are dealing with patients, it was highlighted the need to be careful and to ensure safety. Patient safety must be a priority and requires careful approaches.

''They deal with vulnerable people.. I see how careful they are'' (assessor no. 2)

''The green card is about being exceptionally careful and diligent.. The yellow card is about hurting patients'' (assessor no. 12)

''I would use the yellow card when I think that the student is dangerous'' (assessor no. 14)

Safety is not just about avoiding physical harm during the station. Rather, it is important to be careful to avoid any physical and psychological harm.

''Some of them are arrogant sometimes and think: 'right I can do this', and then they come up with some ridiculous things'' (assessor no. 7)

''Rudeness and aggressive questioning'' (assessor no. 15)

A-16 Task completion

The last student-related factor is the need for task completion. The students are expected to achieve and complete what they are asked to do in the station.

''The things I tend to look for is obviously achieving the task'' (assessor no. 9)

''He did complete the station but you could argue that he did not complete it to an adequate level'' (assessor no. 10)

''There was a completion'' (assessor no. 18)

B- Assessor related behaviours

B-1 Calibration

The first assessor-related factor is the process of checking and adjusting the application of a measuring scheme. This calibration process is between candidates against given standards. The assessors might spend some time, first few assessed students, to familiarise themselves with the marking scheme and mark the candidates performance against such standards.

''It is easier to assess the later students.. There is a little calibration going on'' (assessor no. 12)

''I think there is always.. again.. a degree of calibration.. which sometimes you kind of.. by the middle you calibrated yourself'' (assessor no. 16)

Such calibration could initiate some difficulty with the assessment process and place some pressure on the assessor at the beginning of the exam until they become familiar with the standards.

''It is very difficult when you have the first few people through.. It is quite difficult to gauge where everyone else is going to be.. The most difficult thing is trying to be consistent with your grading throughout, bearing in mind that you do not see the standard person first of all'' (assessor no. 8)

''Sometimes I feel a bit pressurised at the beginning.. examining the first student I think you have an element of level finding to sort of think about what you are expecting from students at that particular stage, and so they do sort of become a baseline'' (assessor no. 10)

Additionally, and besides familiarising oneself with the assessment criteria, the assessors might spend some time familiarising themselves with the station, the questions, and the time required to complete each task. They also familiarise themselves with their roles as assessor when it comes to how and when to prompt.

''I think the first ones (students) are not easy either because you kinda warming up, getting familiar with the station or the questions ... so, I think probably the ones in the middle are better'' (assessor no. 13)

''Definitely the last students are easier to assess because by then you know how the timing flow of the OSCE station, so you know when to prompt and how to make sure the student gets through everything''(assessor no. 15)

It was found that comparisons between students might happen. It is important to note that some assessors could fall into the trap of comparing one student to another, instead of comparing each one against the given standards.

''Comparing one student to another could be an issue because you might think you were not fair with the first student'' (assessor no. 7)

''I know you should not compare one student to another, you compare each one to the criteria. I am aware of that but I think it is human nature and it is hard to eliminate that part of it'' (assessor no. 10)

''Comparing students with each other definitely happens'' (assessor no. 12)

Some assessors compared between the two students they observed and assessed when I interviewed them.

''She was not more effective like the first one was'' (assessor no. 1)

''She is polite and warmer than the first student'' (assessor no. 3)

B-2 Reluctance

The second assessor-related factor is reluctance or their hesitation or uncertainty when it comes to which mark the student should receive. The assessors sometimes can be unsure and they might spend more time in deciding which mark they will award the student. Looking back to Table 2, more than half of the assessors, 11, were reluctant to some extent in their decisions. They first gave two decisions, and then they took some time to go with only one.

One type of reluctance is when it comes to failing a student. Some assessors find it hard to fail a student and therefore they spend more time before they make their final decisions. This is also a part of being lenient as will be discussed later.

''By the time they (students) come to the exam, you know there is a critical point in their training, and therefore you really do not expect in a way anybody to fail. I find it difficult and try to think: is the student doing what I think he is doing? Is he saying what I think he is saying? So I double my effort to understand that the student is failing'' (assessor no. 13)

B-3 Observation skills

Examining a student using the OSCE is about observing their performance. This observation of performance might be different from one assessor to another. Different assessors have different observation skills. Therefore, some assessors might miss some parts of the consultation more than other assessors, and this can be reflected in the marks they give their students.

''I do not consider myself very good at considering the nonverbal communication between two people I am watching. I can consider the nonverbal communication between me and an individual.. It is very difficult.. I imagine a scenario that I could receive training in.. Certainly I am not skilled at the moment'' (assessor no. 4)

B-4 Dove vs hawk

Leniency is another point that could influence an assessor's judgement. Leniency level varies between assessors from dove-like assessor to hawk-like assessor.

''I am a soft examiner. I do not want good people to be throwing five years away because of they had a bit of mental block on the day'' (assessor no. 9)

''I am not nearly as harsh in an actual OSCE though. I am a lot softer when there is a real student in front of me'' (assessor no. 7)

''I tend to be rated towards the dove end but I do not apologise for that'' (assessor no. 9)

It was also found that leniency may be more apparent and noticeable when it comes only to failing a student. The assessor might feel relaxed about awarding the student whatever mark they think the student deserves except when it comes to failing him or her.

''By the time they (students) come to the exam, you know there is a critical point in their training, and therefore you really do not expect in a way anybody to fail. I find it difficult and try to think: is the student doing what I think he is doing? Is he saying what I think he is saying? So I double my effort to understand that the student is failing'' (assessor no. 13)

Leniency was found to be variant according to the year of study the student was in. While some assessors were generally lenient, other assessors were more lenient with earlier

years of study students because there are less expectations with younger students. Some assessors might become less lenient with later years of study students because they do not want to graduate unqualified doctors.

''I tend to be more lenient with students in year 3'' (assessor no. 2)

''Less expectations with younger students'' (assessor no. 8)

''I am more strict with older students'' (assessor no. 11)

''I am less strict with year 3 compared to year 5'' (assessor no. 13)

One type of leniency found was about being more sympathetic whenever the student is very nervous and anxious. This type of leniency may not influence the assessor's judgement in a direct way. Rather, the assessor might somehow encourage the students and give them leeway which could in turn influence the student's performance in a positive way. Although it might be expected to see this kind of leniency more among those assessors who had experienced OSCEs themselves, no noticeable difference was found in this research with this regard. Some of the assessors who had never experienced the OSCE themselves as students did show this kind of leniency.

''They are horribly nervous ... They can be very intimidated'' (assessor no. 1)

''Whether the student might be really nervous .. Sometimes you try to give them a little bit of encouragement even by just a smile. Sometimes I do that'' (assessor no. 11)

''Yes you do try to be objective about it, but there are occasions when subjectivity feelings come in. You can see they are anxious.. you may give them a little bit of leeway'' (assessor no. 3)

Finally, it was interesting to find that it would be possible for some assessors to choose to be lenient only when they are tired or not in a good mood. They think this chosen leniency would eliminate any unfairness caused by being tired or not concentrating during the station.

''I tend to be more lenient with the students when I am tired or in a bad mood. I do not want them to suffer from that'' (assessor no. 1)

B-5 Intonation and accent

Although it would not be very common, a student might not understand one or more of the assessor's questions because of the assessor's accent or intonation. This might happen when the assessor speaks English as a second language regardless of whether the student is a native English speaker or not. It would be an issue if the assessor does not notice that the student did not follow him or her because of their accent.

''Sometimes it is the language (sighed). I think most of them speak good English. I mean I am not a native English speaker. So, sometimes they do have a problem to follow my accent.. It is very rare'' (assessor no. 13)

Furthermore, it is worth mentioning that assessors who speak English as a second language might not be as good as native speakers in assessing verbal communication skills. Therefore, a student who is an excellent communicator might not get the same mark when examined by two assessors, one of whom is a non-native speaker.

''Verbal communication skills is easily assessed by native speakers'' (assessor no. 4)

B- 6 Concentration and boredom

Assessors might miss observing some behaviours or listening to some phrases at the station because of a dip in concentration over a period of time during the day of the examination.

''With video you could look back at it and maybe see things you missed in real time'' (assessor no. 5)

Assessors need to concentrate during each station to make sure that they observe every student's performance optimally. It was found that concentration might decrease after seeing many students either because of tiredness or because of boredom resulting from the assessors observing quite similar performances over a period of time.

''As an examiner, it can be difficult to concentrate on a long morning when you are having people doing lots of different things, basically the same style but in a different way'' (assessor no. 3)

''Perhaps there is more interest when you start.. as you getting on you start to get a bit sort of tired or jaded with hearing the same thing again and again'' (assessor no. 10)

''I suppose I might be subdued if I am exhausted'' (assessor no. 12)

Therefore, the assessors are usually fresher examining the first students while they can be tired and jaded assessing the last few students. Consequently, there might be variances in assessing two similar performances when one assessed first and one assessed last.

''I find assessing the first students easier because you are fresher, less bored'' (assessor no. 14)

''If you are doing the same exam, after ten students it is hard to keep concentration'' (assessor no. 14)

''I think sometimes first (students) is slightly easy because you are fresh, and actually by the end you are slightly jaded'' (assessor no. 16)

Likewise, it was highlighted that breaks are necessary to keep assessors focused during the day of the examination. Such breaks could help the assessors to hear something different and get some refreshments. It was also highlighted that moving from one station to another helps increase attention and concentration.

''I think you do need to move from station to station a bit. I think probably 5 or 6 maybe in one station would be adequate, and then you should shift to a different one to maintain some sort of clarity.''(assessor no. 3)

''I like the breaks during the exam. I think they are important to make sure that we stay focused.. Talk about something different for five minutes''(assessor no. 4)

Finally, and as mentioned earlier, it was highlighted that tiredness might encourage assessors to choose to be more lenient to avoid any unfairness caused by such tiredness.

''I tend to be more lenient with the students when I am tired or in a bad mood. I do not want them to suffer from that'' (assessor no. 1)

B-7 Idiosyncrasy and own standards

It is important to note that the assessors, as with other human beings, have different experience. Such experience could shape the way they think and make decisions.

Likewise, assessing students might vary from one assessor to another in certain ways that depend on personal experience.

First, it was found in this research that assessors' own experience as patients has informed their decisions when they assess students. They might encourage or discourage certain things they liked or disliked when they were patients seen and treated by other doctors.

''I saw bad communication skills when I was a patient, and I am aware of them now'' (assessor no. 1)

''When I was a patient I wanted my doctor to be more receptive'' (assessor no. 2)

''I like them to be thorough, because things were missed in my diagnosis because people were not thorough'' (assessor no. 4)

Second, it was found that the experience of one or more of the assessor's relatives as a patient could inform their decision. The assessor might discourage or encourage a certain behaviour based on a relative's experience.

''The patient sometimes wait for a long time to be seen by a doctor… One of my relatives was annoyed when her doctor was in a hurry'' (assessor no. 2)

''My father was admitted to the hospital, and the doctor did seem a bit arrogant and disinterested'' (assessor no. 3)

''I have seen my parents sufferance, and so I am aware that we can always do a lot more to explain what is going on. You can almost never do enough to explain exactly what is going

on. And so that as an assessor I like to see that explaining, clarifying, checking. I like to see that in students and doctors'' (assessor no. 12)

Third, some assessors would sometimes put their own way of doing a procedure or technique as a standard. So, when the student does it differently, it might influence the assessor's decision.

''I have a huge bias… We have got to be very accepting of a wide variety of which way will people do it.. and that is another problem with the OSCE format because it is time limited'' (assessor no. 4)

''It can be difficult to follow somebody when they do not really know what they are doing.. They do not follow as you may do when you consult'' (assessor no. 3)

''Sometimes the student does not do what I expect him to do'' (assessor no. 2)

Fourth, the assessors might sometimes put themselves in the patient's shoes. Would they, as a patient, like what the student is doing? This might lead the assessor to look at the patient's face to see whether they are happy or annoyed.

''If I were the patient would I understand what the student is saying'' (assessor no. 2)

''I try to put myself in the patient shoes, sometimes not always'' (assessor no. 3)

''I do put myself as the patient, whether I would find that student easy to communicate with.. whether I would like him as my doctor'' (assessor no. 11)

Fifth, some assessors have their own standards that might not match with the given mark sheet. Own standards might make some assessors struggle to compromise. Such own standards could interfere and influence the assessors' decisions.

''I am struggling! Because in my mind I would make a borderline, but actually if I go by the criteria I would give her a pass'' (assessor no. 16)

''Using the descriptors alone, clear fail.. I went straight to borderline and I tried to justify giving this candidate a borderline view. I certainly did not expect to have to go beyond that'' (assessor no. 4)

''I get my own standards by seeing more students'' (assessor no. 14)

''There is always a temptation just to notch that mark sheet up and notch that mark sheet down'' (assessor no. 18)

Sixth, it was found that subjective and personal feelings and preferences, not necessarily true, might play a role in how an assessor makes a judgement. This kind of feeling might cause some bias if it was not controlled adequately when examining a student. More details about bias will be discussed later.

''You just get a feel for somebody you cannot always put it into words.. You get a feel for somebody whether you would like them to be doctors'' (assessor no. 7)

''I think the OSCE is quite subjective, because I think it is very difficult not to get a feel for the person when they come in because you see them face to face. The personality can pop off and it is sometimes more difficult to be objective'' (assessor no. 8)

''We need to criticise the behaviour not the person'' (assessor no. 18)

''If they talk with their hands that kind of things annoys me'' (assessor no. 9)

Seventh, assessors who also teach can have different expectations from the students. Teachers are usually more aware of the curriculum and the teaching and learning

outcomes. Therefore, they might have more expectations than assessors who are not involved in the teaching process.

''I teach ... If the students are examined on the things I gave them, I want them to say it'' (assessor no. 7)

''I generally have examined on students I have taught.. I have relatively a good idea of the level of performance that I would expect from them. So, it is quite easy for me to slip into: this is what I would expect, this is what the sheet is expecting'' (assessor no. 18)

Eighth, some assessors would be more interested than others in checking for the student's understanding even though sometimes it would not be possible to do so during an examination such as the OSCE. This interest in exploring the student's understanding is sometimes beyond their responsibilities as assessors. Such an interest might increase subjectivity and cause some variance between two assessors when they examine one similar performance.

'We do not really test true understanding of the material'' (assessor no. 4)

''It is really very difficult for me to probe whether a candidate does understand in certain circumstances.. My interest is: do I feel the candidate has truly understood the information they have been given'' (assessor no. 4)

''Examiners are discouraged from exploring in greater detail the student's understanding of what is going on'' (assessor no. 12)

''They can percuss a normal chest, but I can get no information really from the structure of that station whether they understand why they are percussion the chest'' (assessment no. 5)

Ninth, observing or talking to colleagues has been found to inform the process of making judgements and decisions. It would be expected for assessors to meet and discuss different things such as what they expect of their students in an OSCE station. It would also be possible that some assessors observe each other in, for example, training programs and exchange knowledge. Such experience was found to be a possible factor in informing assessors' decisions.

''I suppose talking to colleagues does change how you may assess in terms of expectations, so what do you expect from students'' (assessor no. 16)

''Having seen communicators as a student myself... Some of them are appalling communicators'' (assessor no. 8)

Finally, it was found that different assessors might have a different interpretation of one particular characteristic. Measurements such as 'depth', 'amount' 'size' or 'adequacy' might be seen differently from one assessor to another. For instance, the word 'partial' can be interpreted differently and subjectively. Individual interpretation of an action or standard was found to be a possible factor that can influence assessors' decisions.

''There will always be a little variation in interpretation.. For example, did the student take a drug history about something? You would need to make a decision as to whether you are going to give them the full mark or a midway mark if they have done it partially, and that 'partial' I think will be a little bit subjective'' (assessor no. 10)

The next three quotes clarify this idea more clearly. These quotes are what three assessors described regarding the adequacy and appropriateness of one certain part of a student's advice and consultation about the patient's life style.

''She did too much on the social waffly bit about keeping a happy life style and silly comments on that'' (assessor no. 15)

''She gave a reasonable balanced explanation of how to follow a healthy life style from there'' (assessor no. 18)

''Some of the life style part, it was an odd thing to say'' (assessor no. 3)

Such difference in interpretation played a significant role in making judgements. The next two quotes are what two assessors described regarding one behaviour of the female student when she looked at her watch while seeing the simulated patient.

''The reason I don't think it is a very good pass is because she was still occasionally distracted. She looked at her watch twice'' (assessor no. 6)

''You could argue about the watch, it is a very minor thing'' (assessor no. 18)

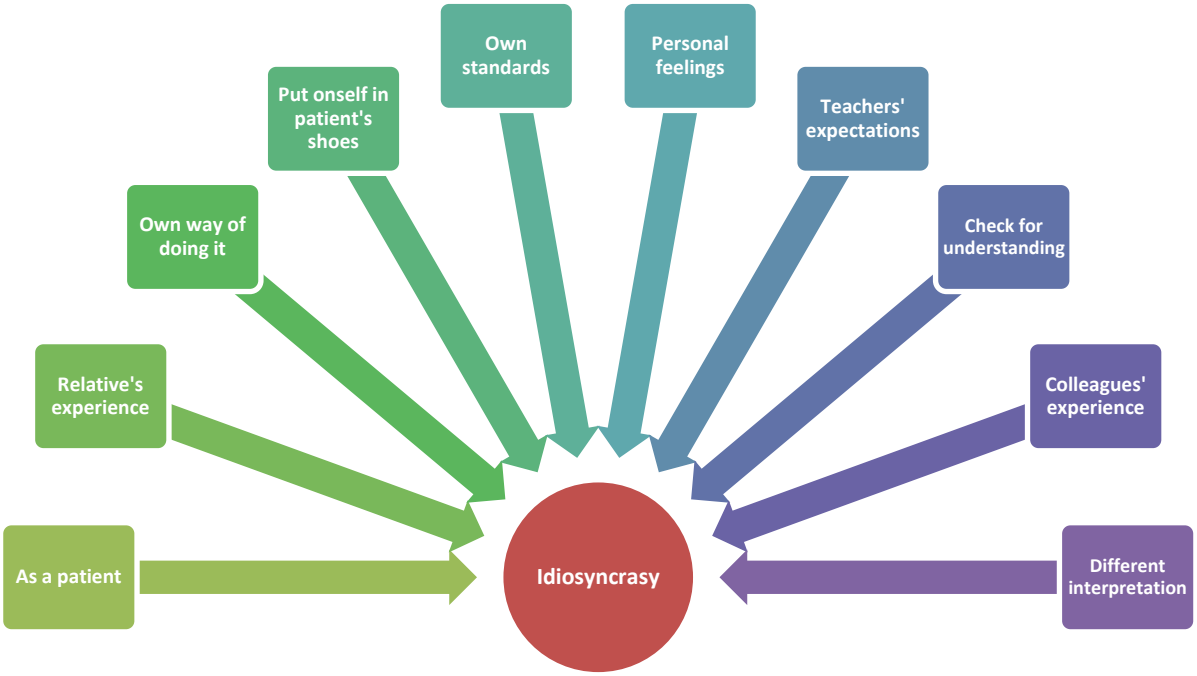

Figure 13 Assessors' idiosyncrasy

B- 8 Self-discipline

Self-discipline was found to cause some pressure and challenge when assessors examine students in the OSCE. Assessors in the OSCE are asked to follow certain rules and regulations in order to increase reliability. For instance, the assessors are not expected to ask extra questions for clarification.

''The main challenge that the OSCE assessors face, I think, is the discipline that it places upon yourself, to only prompt where it is appropriate, to keep an eye on the time'' (assessor no. 9)

''One aspect that perhaps not so useful is the fact that for robust reason of fairness, examiners are discouraged from exploring in greater detail the students understanding of what is going on.. It is both but I think it is more objective than subjective because of those restrictions and restraints'' (assessor no. 12)

In addition, some assessors are more aware of the responsibility and mission they are expected to fulfil in order to produce fair and reliable decisions. Such responsibility could place some pressure on them.

''From the examiner point of view, probably the most frustrating thing about the OSCE is trying.. to maintain a consistency with each candidate that comes through'' (assessor no. 6)

''Sometimes you get tired. It is important that you tell yourself that you actively treat the last one in the session equally'' (assessor no. 9)

''One of the challenges is trying to be very even-handed.'' (assessor no. 18)

''It is important to kinda discipline yourself. You may have heard this five times already this morning, but for the student it is the first one. So, yes you might want to mark him harsher, but do not mark them any more harshly than the first one'' (assessor no. 9)

It was also found that sticking to the marking sheet and scripts might cause a challenge. Some assessors might need to put more effort in to fulfil this requirement and discipline.

''I am required to very much stick to the marking schedule'' (assessor no. 4)

''From the examiner point of view, probably the most frustrating thing about the OSCE is trying to not say anything above and beyond what is in the script'' (assessor no. 6)

B- 9 Seeking patient satisfaction

During the exam, some assessors might tend to look at the patient's face or body to see whether they are satisfied or annoyed. The assessors use such clues and information to help them make or support their decisions about the student's performance and attitude.

''I thought it might be difficult or to realise what the perception from patient's point of view might be of particularly her fiddling with her pen''(assessor no. 10)

''If I see that the patient is uncomfortable, then definitely that will influence my marking'' (assessor no. 13)

''I would mark the student down if I felt that the patient was not happy'' (assessor no. 14)

''I look at the patient during the exam because he or she might be frustrated if the doctor is not listening to what they are saying.. So I think you need to look at both of them (student and patient)'' (assessor no. 15)

''It is important how the patient responded back to them. Say I was looking at the patient and the patient was smiling, the patient was happy..'' (assessor no. 18)

This need to look at the patient's face while examining a student in the OSCE let some assessors comment on the assessment of a student using a video, instead of face-to-face. With the video, as in this research, the assessors did not have the chance to look at the patient's face clearly.

''One issue with the video is that I cannot see the patient's face and how she responds to him.. You will only rely on him (student)… 'If the patient does not show shock or surprise then I will not be worried.. I think as long as the patient is happy I would accept most things'' (assessor no. 3)

''The thing with the difficulty with the video, particularly as it is short like that, is that I cannot see the patient's reaction at all. I cannot see anything about her facial expressions'' (assessor no. 8)

Some assessors in this study were interested in seeing how the patient would mark the students. The judgements of the assessor and the patient need to be completely independent from each other. However, seeking patient's satisfaction might risk this independence in assessment and making decisions about the student's performance.

''We have got the mark from the simulated patient. They are obviously giving assessment for the trainee, and I know that has been available now for a number of years, and obviously you will be interested to see what the correlation between that is and the marks that given by the assessor'' (assessor no. 6)

''The patient did not appear to object, but it would be interesting to see what their impressions were'' (assessor no. 18)

Seeking patient's satisfaction can be more apparent if it is a real not simulated patient. Responses from real patients could be seen as more genuine.

''I think with the actors it does not give a lot of way because they are very much the same, but if it was a real patient who is being examined it does give you some information about how they are feeling, if they are feeling relaxed.. I might do mark the student down if I felt that the patient was uncomfortable for some reasons'' (assessor no. 17)

''I think very much you would look at the patient's responses, particularly that is more important with the real patients than the simulated patients'' (assessor no. 18)

B- 10 Bias and stereotyping

There was some stereotyping going on while the assessors were assessing the two students (a male and a female student). For instance, some assessors commented on some gender differences with regard to confidence and listening.

''May be male students sometimes are not as good at listening to patients.. I am trying to think if I have got any evidence for that basis, or whether that is just a cliché I have come up with'' (assessor no. 17)

''Boys are just more direct and forthright. Girls tend to be a little bit more quieter.. women in the whole tend to listen better'' (assessor no. 5)

''Females tend to be more tentative sometimes. I would say the male students in general tend to be more confident''(assessor no. 8)

Some assessors gave general impressions about all female students, that they are better than male students in several features such as empathy, listening, and eye contact.

''I think there is a difference, and I think it is important to acknowledge the difference.. I find that females, not just students but junior doctors as well, probably seniors are.. they find it much easier to develop empathy, to have eye contact.. My impression is that they seem to have much more innate understanding of nonverbal communication'' (assessor no. 4)

''Women in the whole tend to listen better'' (assessor no. 5)

Another assessor gave a general negative impression about male students that they would be more likely to get a yellow card than female students.

''I have not given anybody the yellow card but I wondered sometimes it would be more in the male students'' (assessor no. 7)

One possible type of bias was found to be about the language and fluency of speech and tone of non-native students when they see native patients.

''Should they (non-native students) be allowed to consult in an exam situation in a non-native tone? I do not know the answer'' (assessor no. 4)

Assessing a student that the assessor has seen before might cause some bias. Recalling previous performances or attitude has the potential to influence the assessor' decision and judgement.

''You cannot eliminate bias if you do know the student.. It is easier and fairer to assess a student you have never seen before'' (assessor no. 12)

''I think if you know a student it is inevitable to have recollection of what the student was like when he or she was in the hospital under your supervision, and sometimes you establish a good relation with the student and then you remember: yes, this student was good or not so good.. It is inevitable in the back of your mind.. Yes it can affect my decision; if it is in the borderline, then I think probably they will pass'' (assessor no. 13)

The way a student enters the station and performs a procedure may play a role in influencing an assessor's judgement. This cognitive bias might be referred to as what is known as 'halo effect' where a general impression of a student can influence the assessor's feelings about that student.

''There is always going to be some bias I think.. For example, I suppose somebody might come across as being very good and efficient of what they are doing, but they might not necessarily cover the actual points that are in the assessment, and therefore you feel they have done really good job but they might not have actually covered everything that is supposed to be covered. So, there is a feel that you want to give them a higher mark when in fact actually if you are just looking at what they have done, it might be the same as somebody else who was not so good.. I am a bit biased with those kind of students. May be I do try and find ways to give them more marks which is maybe not fair'' (assessor no. 17)

Personal and individual preferences and beliefs can play a role in how an assessor makes a decision. Individual likes and dislikes might affect how someone feels about another person. Likewise, the OSCE assessors can have a feeling about their students which sometimes can cause bias in judgement as such feelings are not related to academic achievements and performances.

''Everybody has got prejudice, not necessarily in a nasty fashion, but things that would annoy you, things that you would like. We are used to that with patients, so all the time with patients you are aware that this patient has this particular issue problem, akhh, they really annoy me patients like that, but I can't let that alter or interfere with what I do. So, most doctors are able to take that through the same thing with OSCEs. They may have prejudices, they have likes.. dislikes..''(assessor no. 18)

''We need to criticise the behaviour not criticise the person.. You never comment on what you believe the students actions are, you only comment on what you see'' (assessor no. 18)

B- 11 Assessor confidence

Confidence was identified to play a role in assessors' judgement. Confident assessors found the process of assessment easier compared to less confident assessors. It was found that two elements could influence confidence. First, being familiar with the OSCE as an assessment method and the required process and procedures was a factor that helps increase confidence.

''I was quiet nervous when I first started assessing them'' (assessor no. 8)

''Initially I was of an age where we did not have OSCEs coming through. So, when I first came to OSCEs, completely fresh to me. There was not any form of examination I had done before.. One of the things when I first started actually doing the OSCEs that I found most difficult was, as an assessor, I was paying more attention to the actual marking sheet and making sure that did not forget anything rather than necessarily paying much attention to what the student was doing.. But after a couple of OSCEs you can manage the organisation side quite well'' (assessor no.18)

Second, being familiar with the content of the exam and the case itself was another factor that could help increase confidence.

''And each time I have done a station it is being something that I felt comfortable with examining on in general practice. It has not been anything that I felt it was outside my sphere of knowledge''(assessor no. 8)

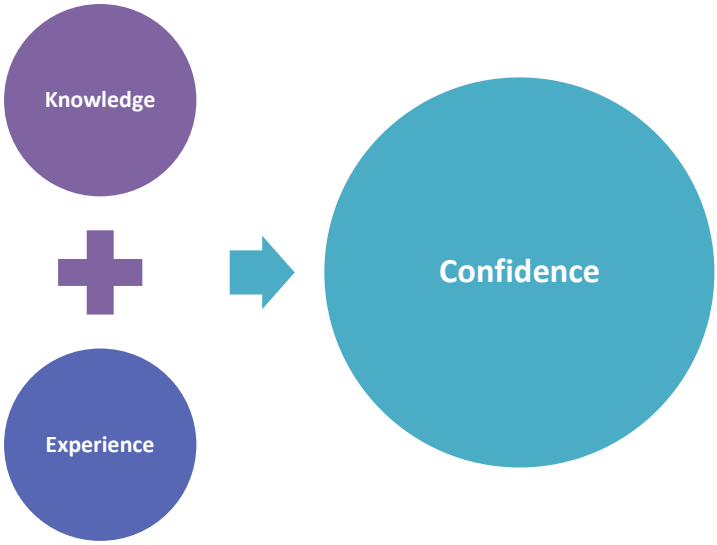

Figure 14 Assessors' confidence

B- 12 Recall

It is expected from the students in the OSCE to perform several things during the allocated time to each station.

''One weakness of the OSCE is that students will be doing so many things in a very short period of time'' (assessor no. 2)

Therefore, it is very possible that some assessors forget what the student did perform and what they did not. As a result the judgement of performance can be inaccurate.

''It can be difficult to remember everything they have done'' (assessor no. 3)

''You are very much reliant upon your memory and the little notes you make during the consultation when you sat there as an assessor. So, I guess if you are looking at accuracy, it probably would be more accurate to assess a video.. You can rewind''(assessor no. 18)

C- Patient-related behaviours

This section highlights the influence simulated or real patients might have on both candidates and assessors during an OSCE station. Although both real and simulated patients were found to have an influence, simulated patients were considered to be more effective, as will be described later.

''Simulated patients are generally very good'' (assessor no.4)

''Most of simulated patients are very good'' (assessor no. 5)

''Simulated patients can be very good'' (assessor no. 16)

C-1 Consistency

Consistency in performance is required to ensure that each student receives the same experience. This helps increase fairness. However, some issues with consistency exist.

''It is very very hard even for the same simulated patient to give the same performance and act in the same way to each student'' (assessor no. 18)

Consistency differs between real and simulated patients though. It was highlighted that simulated patients are more reliable and easier to get information from, compared to real patients.

''In term of delivering the history, simulated patients are more consistent and more reliable.. I think there is an increasing move towards not using real patients unless you need to demonstrate a physical sign'' (assessor n. 9)

''Most of them (simulated patients) are very good.. Patients on the other hand do not always give the same story, and they sometimes forget things in certain times'' (assessor no.5)

Boredom might cause inconsistency. It is possible that some patients get bored after spending a long time doing the same thing with every student. This boredom could affect their performance, reactions and therefore consistency.

''They may get bored when they do the same thing several times on a morning'' (assessor no. 3)

Adherence to the transcript is essential to maintain consistency. It was identified that some patients could add more or even unrelated information while consulting a candidate. This could risk fairness and replication.

''I have probably been irritated by one or two of the simulated patients.. I feel they should stick to their part'' (assessor no. 3)

''Sometimes some of the simulated patients fluff their lines''(assessor no. 6)

''I think it is much easier to examine people on simulated patients because they do not talk about completely ridiculous things.. it is not really authentic but they make it easier.'' (assessor no.7)

''Generally if it is an actor simulating a patient I think that may be a little bit more objective because they stick to the script a little bit more'' (assessor no. 15)

Replication can also be easily influenced when the patients feel tired and fatigued after a period of time. Such tiredness might be more clearly seen among real patients than simulated patients.

''I think it is very tiring experience for the real patients, particularly the type of patients who are available to spend a whole day in the medical school'' (assessor no. 9)

C-2 Language barriers

Students who speak English as a second language might find it difficult sometimes to understand a slang word or a phrase said by a simulated or real patient. In addition, the intonation of the patient might change the meaning of a word. Such misunderstanding might cause inaccuracy in what the student would do next or in how they would answer the patient. This issue might be seen among real patients more than the trained simulated patients.

''If it is a real patient there might be language barriers because they might use terms a foreign student would not be aware of'' (assessor no. 15)

''If English.. it clearly is not their (students) first language, then sometimes patients or the simulated patients can say things and they do not always necessarily pick up or understand

the phrase that can be used. But usually simulated patients are good enough so they can rephrase it'' (assessor no. 5)

C- 3 Dove vs hawk

It was highlighted that some patients can be harsh or generous in how they deal with a student. This can be seen in their satisfaction and in how they cooperate with the student in the station.

''Who are either deliberately obstructive to the student, or are far too generous..'' (assessor no. 3)

''Some patients can always be dissatisfied. So it would depend if the patient dissatisfaction was justified.. You cannot make everybody happy'' (assessor no. 12)

''I think the simulated patients are fine, just I think that some of them are maybe a little harsh in marks than others'' (assessor no. 7)

Furthermore, discrimination and racism might happen which could make it difficult for some students to perform in an optimal way.

''Some of our patients can be quite racist. You get an elderly person who got their fixed ideas'' (assessor no. 7)

''Simulated patients will not affect my assessment, but I think if it is a real patient that could be a problem definitely.. The real patients could give them a hard time.. They could be very rude to some students from different cultures'' (assessor no. 7)

C-4 Culture-related

A student who comes from a different culture might not be familiar with some culture-related behaviours that a patient might display and expect the student to respond to. Such inability to respond to certain behaviours might be interpreted differently and could influence both the assessor and the patient's judgements.

''It is an interplay between patient and doctor. The way some patients behave and the way they want response to them is hard to pick up if you are not used to picking them culturally'' (assessor no. 14)

''Some of the body language maybe different in different cultures, and that can be perceived differently'' (assessor no. 16)

C-5 Adaptation

Different patients have different abilities to adapt their behaviours and responses according to different situations when seen by a student.

''The difficulty with the real patients is that they can't vary their response depending to the student approach. So, they only can be themselves'' (assessor no. 18)

''Some simulated patients are very good, some of them are rather too self-oriented; they hear from the OSCE not the OSCErs (they do not stick to their roles).. They are more interested in themselves than their role within the process'' (assessor no. 18)

D- Environment and organisation

D-1 Preparation

Excellent preparation helps everyone in the station, student, assessor and patient, to concentrate and feel comfortable. It impresses everyone when everything seems in place. Regular breaks will also help the assessors to focus and increase their concentration.

''For me it is often the external things which I find most frustrating rather than necessarily the internal things'' (assessor no. 5)

''It is organised, prepared for us already to let us concentrate on assessing and marking the students'' (assessor no. 13)

''I like the breaks during the exam. I think they are important to make sure that we stay focused'' (assessor no. 4)

D-2 Timing

The time dedicated to each station was seen by many assessors as short. This shortness in time might cause some difficulty for both the student and the assessor.

''It is difficult to assess communication skills in five minutes'' (assessor no. 1)

''It is fairly short time really. You know in a hospital setting you will not necessary be limited to that amount of time'' (assessor no. 3)

''Sometimes I do have the feeling that OSCE is not as objective as it could be depending on the station and also depending on the timing of the OSCE'' (assessor no. 15)

Some assessors might feel pressured as they are asked to examine and mark the student in such a short time before another student comes in.

''Well, the face to face is a real time. With this one (video) I can probably think twice or three times, I take my time, unless you put me in the pressure and say mark them now'' (assessor no. 13)

Furthermore, such shortness in time might place some pressure on the student as they are required to do many things and integrate their skills in a quite short period of time.

''The time limit can cause some stress'' (assessor no. 3)

''The challenge is trying to integrate different skills and do that in a way that can be examined in 8 minutes'' (assessor no. 16)

''They (students) have a time pressure'' (assessor no. 17)

D- 3 Task preparation

The case and task could influence both the student and the assessor. Poorly structured questions and tasks might have a negative impact. It is very important to have well written scenarios and questions. In addition, real and simulated patients need to be fully aware of their roles and tasks.

''Sometimes I do have the feeling that OSCE is not as objective as it could be depending on the station and also depending on the timing of the OSCE'' (assessor no. 15)

''Last year I did not start often in a bad mood, but as the session went on I seemed to be getting in a worse mood, and it was actually about the exam. I got a station where I thought

a very poorly structured question, a poorly structured task, in an area in which I have expertise''

D- 4 The mark sheet

Very broad and less clear marking schemes might increase subjectivity. It was highlighted, by some assessors, that a clear and more specific scheme would be more helpful for the assessor to make his or her judgements.

''I think if the mark scheme is written well it is more objective. There is obviously always that chance of subjectivity.. If the marking scheme is quite broad, the examiner is left thinking not quite sure which way I should mark'' (assessor no. 16)

''It can be difficult to understand from the mark sheet though how much emphasis is being placed on each element of it'' (assessor no. 3)

Doing an overall assessment of the candidate was also appreciated and seen as an additional and helpful procedure.

''It is objective because the marking scheme is very clear, and the one in Leeds does give the chance to do an overall assessment of the candidate''(assessor no. 15)

Well written marking schemes could help minimise any negative impact of tiredness or the process of calibration an assessor might confront while assessing the first few students who come through.

''I think there is always again a degree of calibration which sometimes you kind of by the middle you calibrated yourself.. I think with the objective scheme that should be less of an issue'' (assessor no. 16)

''A good mark sheet can help minimise any effect when I am tired'' (assessor no. 2)

D- 5 Background noises

It is essential to make sure that everyone in the station is concentrating on their tasks. Background noise could easily distract everybody's attention.

''The only two things I have come across or I found that the assessment has been difficult, one occasion was when we got four OSCE stations in the same room. We are only separated by barriers, and you could very easily hear what was going on in the other stations, and that made it difficult to be certain that you picked up everything that the candidate was saying. I found that very frustrating'' (assessor no. 6)

''Sometimes you hear people next door which is not good. I think that is very off-putting'' (assessor no. 17)

D- 6 Temperature

Finally, it is important that the room temperature is ideal to help stop any decrease in attention and concentration.

''There was one day when it was about 30c outside, and it was really hot. It was hot both for the candidates, the assessors and the actors and actresses. I suppose that probably makes it more difficult to make sure that your concentration is fully on the assessment'' (assessor no. 6)

Conclusion

Four main themes were identified to have an influence on assessors' judgements. These themes are related to the three characters in the OSCE, student, assessor and patient, and

the environment in which it is situated. Under each main theme or category, several subcategories were identified. For instance, student-related factors included appearance, confidence, and cultural-related behaviours. Assessor-related bahaviours included calibration, reluctance, and observational skills. Patient-related behaviours included consistency, cultural-related behaviours, and adaptation. Finally, factors such as temperature or background noises can have a negative influence on all the three characters in the OSCE.

Chapter 5 Discussion

In this final chapter I will review the key themes of this research, Figures 6 & 7, drawing them together under the non-verbal behaviours of the three 'characters' in the OSCE and the environment in which it is situated. The results support, describe and add to what the literature says, as will be discussed. Suggestions, based on the findings of this research, will also be made to hopefully help understand issues that influence inter-rater reliability. Each main theme (student, assessor, patient and organisation) will be discussed with their subthemes. Before proceeding to the main themes, this chapter will look at the differences in judgements and triangulate my research findings with key literature. This chapter will also conclude with a focus on the contribution this work makes to the literature.

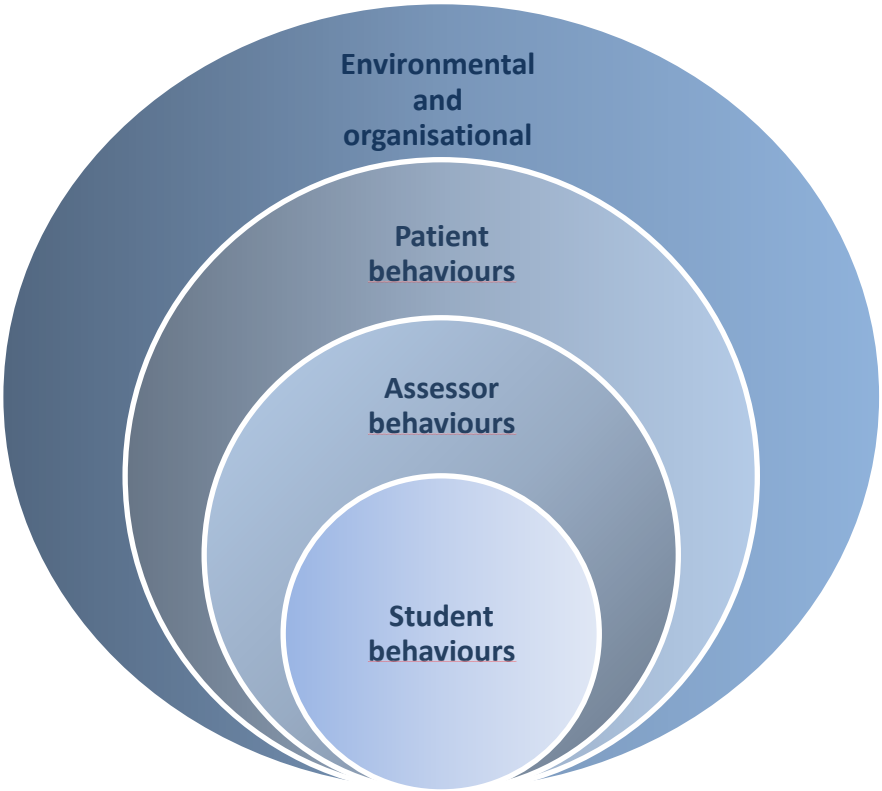

Figure 6 Main themes

Figure 7 Main themes and their subcategories

Differences in judgement

Having watched the two video clips, the assessors formed unique and individual judgements that resulted in varying and disparate decisions regardless of observing the same students and performances. The majority of the assessors in this research made it clear that it was not surprising for such differences in judgements to occur. They clearly emphasised that there will always be a subjective element in the OSCE, and this subjectivity varies from one assessor to another.

Such disparity in making judgements and assessing candidates, found in this research as shown in Table 8, supports what exists in the literature. The literature clearly states that such disparity is seen in assessments that utilise direct observation of performance, such as the OSCE, and assessors' marks can be highly variable in such assessments. Inter-assessor disparities, in different settings, accounted for between 18 % (Alves de Lima et al., 2011) and 21 % (Margolis et al., 2006; Wilkinson et al., 2008) of total score inconsistency-growing to 40 % in one study (Weller et al., 2009). Observing the same performance in this research did not necessarily result in forming the same judgement.

Table 8 Assessors' global judgements and impressions of candidates' performances

Assessor	Gender	Global judgment Student 1	Impression	Global judgment Student 2	Impression
Assessor no. 1	Male	Clear fail	Disinterested**	V. good pass	Excellent
Assessor no. 2	Male	Clear fail	Appalling	Borderline	Receptive
Assessor no. 3	Male	Borderline *	Casual	Borderline	Ineffectual
Assessor no. 4	Male	Clear fail	Rude	Clear pass	Kind
Assessor no. 5	Female	Borderline *	Disinterested**	Clear pass	Sympathetic
Assessor no. 6	Male	Borderline	Disinterested**	Clear pass	Professioned***
Assessor no. 7	Female	Clear fail	Uncaring	Borderline	Pleasant
Assessor no. 8	Female	Borderline *	Arrogant	V. good pass	Empathetic
Assessor no. 9	Male	Clear fail *	Disinterested**	V. good pass	Empathetic
Assessor no. 10	Female	Borderline *	Uninterested**	V. good pass *	Engaged
Assessor no. 11	Female	Clear fail *	Poor.communic-	Clear pass *	Personable
Assessor no. 12	Male	Clear fail	Unprofessional	Clear pass *	Supportive
Assessor no. 13	Male	Clear fail *	Unpleasant	Excellent *	Professional
Assessor no. 14	Male	Clear fail *	Poor	Clear pass *	Fine
Assessor no. 15	Female	Borderline	Unprofessional	Clear pass	Competent
Assessor no. 16	Male	Clear fail	Poor	Clear pass	Average
Assessor no. 17	Female	Clear fail *	Unprofessional	V. good pass	Open
Assessor no. 18	Male	Borderline *	Unaware	Clear pass	Smiley

* The assessors gave two decisions (e.g. between borderline and pass) before they decided to go with only one. ** The student showed 'a lack of interest'. *** ''By professioned I mean came and did the job, but there really was not much extra to it, but was at the line you might expect for somebody at their level.''

The assessors in this research formed different impressions of each student they observed. It has been well established that different assessors will often form different impressions of the same learner even when given the exact same information (Kenny, 1994; Park et

al., 1994). The assessors, as social perceivers, generated in this research different *'person models'* explaining and justifying each model based on what they observed and interpreted. For example, one assessor described the first student as rude because he did not maintain eye contact with the simulated patient. Another assessor described the same student as casual. The literature has discussed this matter declaring that impression formation has been conceived as a procedure whereby perceivers generate *'person models'* of other individuals, explaining what the person is like and why as neurocognitive short cuts (Park, 1986; Park et al., 1994). When the assessors in this research were asked to write down the student's characteristics, some of them did not just list traits. Rather, they connected some traits with characters. For instance, not maintaining eye contact was interpreted as rudeness by some assessors while other assessors justified it as the student being distracted. Some assessors in this research went beyond listing personality traits that explain a candidate by integrating underlying explanations as to why the candidate behaves the way they do or possesses the particular traits. According to Fiske (1993, p. 170), "faced with surprising combinations for which they do not possess ready-made structures, people create brief stories that provide enabling and temporal links among otherwise puzzling bits of information". It has been identified that the '*person model*' shares several features with theories that emphasise the use of social categories as a means to interpret and integrate information about a candidate (Fiske, 1993; Kunda & Thagard, 1996; Skowronski & Carlston, 1989). The suggested reason for having multiple stories for each candidate relies on different combinations and prioritisation of the pieces of information by assessors (Park et al., 1994). Interestingly, in this research not only multiple stories were found, but some stories were contrasting. When one assessor in this study says 'casual' and another

assessor says ‘rude’ to describe one candidate, it shows some level of contrast. Similarly, the female student was described by one assessor as ‘excellent’ and by another assessor as ‘average’. Yeates et al. (2015) described the contrast effect between candidates. However, some level of contrast effect might exist between candidates and assessors. Such variance could ultimately influence assessment reliability and is frequently described as noise resulting from the idiosyncrasy of the social perceiver/assessor (Mohr & Kenny, 2006). Several research studies attempting to document agreement in personality judgements have instead found that these judgements are more frequently unique than similar, even when assessors are presented with the same information about a candidate (Kenny, 1994; Park & Judd, 1989).

One piece of information was enough for some assessors in this research to construct and form an impression of a student. Being very nice with the patient, as the second student showed, or not maintaining eye contact as the first student portrayed, was a main point some assessors used to form their impressions of the students. Park et al. (1994) claimed that perceivers, when forming a model, would attend to a certain characteristic and construct an impression around that central notion. This could explain why different assessors in this research produced contrasting judgements as they may have attended to contrasting characteristics and then constructed contrasting impressions. However, and regardless of the possibility of having idiosyncratic categorisation, the assessors in this research tended to constantly make one of a few possible interpretations of each student. Assessors’ unique way of translating techniques can cause errors in assessment systems that require ordinal or interval ratings when assessors form categorical judgements. Nevertheless, assessors tend to constantly make one of a few possible interpretations of

each candidate (Gingerich et al., 2011). Therefore, person perception is found to be both idiosyncratic and consensual.

Impact summary

It was not surprising to find assessors in this research giving varying judgements and decisions and forming different impressions of the two students. On the contrary, this research aims to understand some of the reasons, non-verbal behaviours, behind such disparity to help understand inter-rater reliability. However, it was interesting to see some assessors give contrasting global judgements and impressions regardless of observing one similar candidate. During an OSCE station, assessors, after ticking boxes, would give their global judgement without explaining their reasons behind such overall judgement. It might be worth asking assessors to write down a few sentences after they finish ticking boxes to explain or justify their overall judgements.

The next part of this chapter will discuss the main themes found: student-related behaviours, assessor-related behaviours, patient-related behaviours and, finally, organisational and environmental factors that can influence all the three previous characters in the OSCE.

A- Student-related behaviours

The OSCE was introduced and designed as a novel assessment method, which could help the assessment of candidates' clinical skills, attitudes, problem-solving and application of knowledge in one examination (Harden et al., 1975). Such a method can help assess students' clinical competence by observing their skills. The assessors in this research

highlighted the necessity of examining such skills, and they used the three types of schemas, as social perceivers, discussed earlier in the literature review (Pennington, 2000). They looked at different sets of behaviours anticipated of a medical student, role schema, such as demonstrating confidence, empathy, professionalism in both style and stress, applying an open approach, and reaching a diagnosis and treatment plan. The assessors in this study also examined what is generally anticipated from candidates' behaviours in the OSCE, event or task schemas, related to the expected sequence of events in such a situation. This might include effects of candidate's behaviours on patient behaviour and organised sequencing. The assessors also made inferences, person schemas, about a candidate on the basis of incomplete available information, through verbal and non-verbal interactional cues in their behaviour. Person schemas contained anticipated patterns of behaviour, personality traits and other inferences about a candidate's knowledge base.

The assessors in this research examined what professional competence (Epstein & Hundert, 2002) includes such as the accustomed and careful usage of skills, knowledge and emotions. Such skills were assessed by observing the two students communicating with the simulated patient. Therefore, communication was an essential means that allowed the assessors to observe, infer, interpret and make a decision about a candidate's performance. It is known that adequate communication skills are required to develop effective physician-patient relationships (Hall et al., 2004). Non-verbal communication can help establish such a relationship through conveying intimacy and interest (DiMatteo et al., 1980; Griffith et al., 2003; Larsen & Smith, 1981). The assessors in this research highlighted the importance of looking at bedside manner and establishing rapport with the

patient through different non-verbal communications and behaviours such as showing respect, listening, body language and eye contact.

Many patient scenarios focus on the candidate's ability to gather history and relevant information from the patient (Tamblyn & Barrows, 1999). Such ability was seen by some assessors in this research through how candidates question, elicit concerns, and provide information. Some of these skills in data gathering and communication might not be available on the mark sheet though. As a result, assessors' global marking might be influenced by one or more of these skills. It was also found in this research that the culture of the student might also play a role in how they question and gather data. Speaking English as a second language or not being familiar with some cultural-related behaviours might hinder the process of data collection. The student language might not be clear and fluent which affects the process of sending and receiving information. The patients themselves may misunderstand the student's questions hence provide inaccurate responses. All of this was found to implicitly influence the assessor's judgement.

Unlike traditional examinations, the OSCE is based on stations that enable contextualisation of competence. It is known that competence should not be seen as an achievement, rather it is a habit of lifelong learning (Leach, 2002). It is contextual and reflects the relationship between a candidate's skills and abilities and what he or she is required to perform in a particular situation (Klass, 2000). Therefore, it was ideal that some assessors in this research highlighted the necessity of observing adaptation and flexibility of candidates. This adaptation is indeed important because health care has been increasingly complex which necessitates and requires conceptualisations of competence

as collective, situated and dynamically produced through social interaction (Lingard, 2012). Competence is assessed to provide insight about the capability to adapt to change, locate and generate new knowledge, and develop overall performance (Epstein, 2007; Fraser & Greenhalgh, 2001). The movement from traditional assessment technique, which pictures learning as planned and formal events with well-defined and unchanging learning outcomes (Bleakley, 2010), to new assessment approaches and strategies helps increase meaning to our assessments. Therefore, in order to assess the complex and multidimensional construct of professional competence, it was reasonable to see some assessors in this research highlighting the important of assessing candidates' ability to adjust and to flexibly apply and develop knowledge and skills when confronting evolving circumstances. In addition, some assessors in this study examined cognitive skills such as problem solving and clinical reasoning which are not considered to be generic (Epstein & Hundert, 2002; Norman, 2003). Such cognitive skills in a specific problem area do not necessarily tell much about the performance of the candidate in other problem areas. Consequently, the context and the task help contextualise competence and assess how candidates perform accordingly.

Candidates, as mentioned earlier, are observed communicating with patients in order to contextualise competence. As a result, it was found in this research that some assessors highlighted the importance of observing how candidates involve the patient in each station. They were interested to see how, for example, diagnosis and treatment plan are shared with the patient. Such involvement can mirror professionalism and mutual respect that can only be assessed through direct observation of candidates. It is known that patient-centeredness and professionalism (Delandshere & Petrosky, 1998; Kuper et al., 2007) need to be inferred from observable demonstrations. In addition, physicians deal

with patients who require practitioners to be responsible for providing optimal treatment and care. "Competence is viewed not only as the possession of knowledge, skills, and attitudes, but rather as the ability to use these in the clinical environment to effect desired results for patients" (ten Cate et al., 2010, p. 674). Therefore, some assessors in this research highlighted the achieved results and impact on the patient. This included safety assurance and completing the task. Concentration during the station is one way that could help assure safety and show dedication. Not paying attention, by being distracted, is seen as annoying not only by the assessors, but by the patients as well. The student is expected to pay attention as a part of adequate engagement. Therefore, whenever the students in the two videos were distracted, the assessors commented on the necessity of concentration in order to ensure adequate engagement and to reach an accurate diagnosis and treatment plan. For instance, and as discussed previously, one of the assessors commented on the first video: *''he was playing with his pen'' (assessor no. 1).* Another assessor said: *"she looked at her watch twice.. She looked at her phone once'' (assessor no. 6).*

Research findings in medical education show that context largely influences behaviours. It is well known that the OSCE is different from work-based assessment. While the latter happens in the real world, the OSCE occurs in a simulated environment (Harden et al., 2015). Such a simulated environment requires role playing. It was interesting to find in this research that some students could struggle, according to the assessors in this research, with the concept of 'role play'. Such a struggle could influence their general performance and therefore assessors' judgements. The culture of the student was seen in this study as a possible reason for such struggle. Some assessors mentioned that the struggle was seen more among international students. As mentioned earlier, meaningful assessment requires assessment being acceptable to both assessors and students. It would be interesting to

explore why such an issue is more apparent among international students. However, it is ideal for the OSCE to be introduced at an early stage. Constructive alignment is an essential component of meaningful assessment, as discussed earlier. In order to achieve constructive alignment, the final examination needs to be aligned to the learning outcomes and teaching and learning activities. Using the OSCE as a learning technique through utilising it as a formative assessment method could largely enable all candidates to be well prepared for high stakes examinations. During formative assessments, the students could share with their tutors and examiners their thoughts about the OSCE. They can ask questions about any unclear ideas or concerns as well as the feedback they receive that could steer their learning.

Finally, since assessors in general can be expected to lack a clearly defined mental representation of the assessment criterion (Yeates et al., 2012), they vary in the way they explain the elements of their anticipation. Furthermore, when assessors observe candidates, the three schemas, discussed earlier, together are used interactively to guide the focus of their attention. In OSCEs, assessors are expected to be aware of the role and event schemas with some slight differences with regard to how much is expected from each student as found in this research. It is anticipated that person schemas will play a larger role in judgement differences.

Impact summary

In this research I wanted to ensure that the assessors saw different skills and non-verbal behaviours when they examined the two medical students. Different skills were highlighted such as adaptation, fluency in speech and performance, and questioning styles. Assessors being idiosyncratic means that different assessors might highlight

different characteristics and then, as mentioned earlier, they would use such observed information to build their judgements. This could ultimately result in varying judgements. The mark sheet that can be used in any OSCE station cannot always cover all areas and skills that some assessors might want to examine. Therefore, some conflicting judgements might arise. In order to help solve this issue, it could be helpful to give assessors some space after they tick all boxes to describe what things went wrong that need consideration. They could explain, in a few sentences, their global marking and the reasons behind such a decision.

B- Assessor-related behaviours

In the OSCE, candidates are observed and scored against a measuring scheme as they rotate around a series of stations according to a set plan (Harden et al., 2015). This raised the issue of calibration between the interviewed assessors in this research and how calibration occurs and differs from one assessor to another. Some of the assessors in this study highlighted the difficulty they might experience when familiarising oneself with the marking scheme, their roles as assessors, and the case itself. The first few examined students might suffer from such calibration. Such effects of calibration on reliability might not be entirely resolved, but could be decreased by ensuring that the assessors receive enough description about the standards, their roles as assessors, and enough description about the case itself.

The assessors in this research used various frames of reference. This study confirms that some assessors used themselves as a frame of reference. This supports what was found in the literature as assessors might use themselves as a frame of reference as they commonly

use their own skills as comparators (Kogan et al., 2010,2011). Inter-rater reliability will be influenced because differences in assessors' experience and clinical skills will ultimately lead to a clear deficiency and variance among assessors when they observe and assess students (Braddock et al., 1997; Paauw et al., 1995; Ramsey & Wenrich 1993; Vukanovic-Criley et al., 2006).

In addition, it was interesting to find in this research that some assessors used their relatives' experiences as patients, the experience of colleagues and the patient as frames of reference. Some assessors in this study clearly stated that they would look at the patient's face to see whether they are happy with the consultation. They believed that the patient's facial expressions can help them gain extra information about the candidates' performance and attitude. However, this research also found, as will be discussed later, that simulated and real patients can be inconsistent because of fatigue or boredom. This inconsistency would discourage building judgements based on patients' facial expressions. Such judgements might be accurate sometimes, but it can also be misleading and inaccurate. The 'assessor as trainable' perspective (Gingerich et al., 2014), as discussed earlier in the literature, refers to either the assessor applying assessment criteria incorrectly, using varied frames of reference or making unjustified inferences which lead to variance among assessors. This study explained this perspective adding new examples and descriptions of how such perspective occur and influence reliability.

During or even after the process of calibration, it was found in this research that the rotation of candidates ultimately can cause some assessors to compare between students. The achievement differences between and among candidates may be highlighted to produce a dependable rank order of candidates across a continuum of achievement from high achievers to low achievers (Stiggins, 1994). This could suggest that some assessors

might compare between candidates because it is easier than assessing them against given criteria. Two of the assessors in this research compared between the two students they observed. This could raise the issue of assimilation and contrast effects. One source of variability is known as criterion uncertainty which means that assessors' criteria are uncertain, constructed differently, or influenced by recent exemplars (Yeates et al., 2013a). The latter refers to the influence of providing a reference point, or an anchor, on judgements made on subsequent problems, or a target, which has been thoroughly investigated by the psychology literature (ibid). However, this research does not confirm any of these two effects as it was not intended to investigate such an effect. This study confirms that comparison between candidates in general occurred by some assessors. The need for greater clarity about the connection between the assessment and what it represents led, in the early 1960s, to the development of what is known as criterion-referenced assessments (William, 2000) for achieving and securing meaningful assessment as discussed earlier. However, a question can be raised: have we reached a limit of criterion usage? As discussed earlier, checklists cannot always cover everything seen and observed during an OSCE station. Allowing assessors to articulate what is not listed in the marking sheet could help justify and understand differences in judgement among assessors.

In this study, eleven (61%) assessors were reluctant to make a final decision about one or both of the two students observed. Whenever they did not feel confident making a decision, the assessors in this study gave two decisions before they were asked to make a final decision. People meta-cognitively evaluate the suitability of their own decisions, and feel confident when this evaluation indicates that the judgement will possibly be correct (Koriat, 1993; Mitchum & Kelley, 2010). Such confidence was found in this research to

increase whenever the assessor was familiar with the OSCE process and with the case presented and task asked. Decision confidence reflects metacognitive inferences by an individual about their adequate ability to make the judgement (Mitchum & Kelley, 2010). Such ability might decrease as a result of assessors not being able to articulate what is not listed in the marking sheet. This could cause some conflicting judgements where assessors might spend more time compromising before they make their final judgements and decisions.

In addition, part of this reluctance was found to occur because of the fact that some assessors can be more or less lenient than others. Harasym et al. (2008) investigated assessment approaches in undergraduate family medicine objective structured clinical examinations (OSCEs) and found that eliminating hawkish (stringent) and dove-ish (lenient) influences changed the outcome for around 11% of learners. In a different study which included 2000 assessors (McManus et al., 2013), around 2% of them were statistically significant hawks and 2% significant doves. While some assessors in this research stated that they generally tend to be lenient, other assessors showed leniency only when it was about failing a student. Tweed and Ingham's (2010) study revealed that assessors were mostly over-confident in their decisions around the threshold of adequate performance, but were under-confident at extremes of performance. Furthermore, the year of study the student was in was found in this research to increase or decrease leniency. Some assessors in this study declared that assessing students in year three, for example, is different from assessing final year students. They justified this trend by saying that the former group is still learning and performance of students clearly changes during learning. However, this raises the issue about the level of expectations assessors have about candidates.

The level of expectations varied among assessors in this research. For instance, some of the interviewed assessors were tutors and they gave different levels of expectations because they would expect to observe what they teach. This fits well with David Boud's (2000) concept of 'double duty' where some assessors may be more able to assess because they teach, but this increases the chance of them being conflicted – what are they assessing? Students' growth? Development? How well they know the taught subjects? Or competence? Furthermore, the literature discusses that noticeable disparity exists in assessors' perceptions of the level at which learners typically perform. These perceptions served the assessors as a general criterion, and were experientially derived, differently constructed, and frequently unclear (Yeates et al., 2013a). Hence, there were differences in comprehending and using the assessment criteria among assessors in this research. As mentioned earlier, assessors in the OSCE are expected to be aware of the role and event or task schemas with some slight differences with regard to how much is expected from each student. Therefore, it is anticipated that person schemas will play a greater role in judgement differences. One way to decrease the influence of person schema on inter-reliability is to ensure that all assessors have clear expectations of what candidates, at different years of study, are required to perform and at what level.

It was found in this research that concentration, fatigue, memory and some issues with observation skills can influence reliability. Cognitive and social psychology affirm that assessors cannot perfectly observe and capture performances (Ilgen et al., 1993) as human memory and processing capacity are imperfect (Baddeley, 1994). Gingerich et al. (2014) justified this issue with reliability by seeing the assessor as 'fallible'. Some assessors in this research stated that they do not want the candidates to suffer from such factors that are unrelated to their clinical competence. They clearly stated that they might be more

reluctant or even lenient when they make their decisions. In addition, some assessors in this study mentioned that they are aware of such factors that could influence their reliability, and this could increase what they called ‘self-discipline’. It was found in this research that self-discipline can cause some pressure and challenge when assessors examine students in the OSCE. This of course varies from one assessor to another, and therefore the effect can also vary.

Bias was also found in this study to influence reliability. Supporting what Yeates et al. (2013b) found, some assessors in this research declared an awareness of different biases when assessing candidates. Bias about the language and fluency of speech and tone of non-native students when they see native patients was identified in this research. A statement like *’’Should they (non-native students) be allowed to consult in an exam situation in a non-native tone? I do not know the answer’’ (assessor no. 4)* shows some level of ethnicity or cultural bias. Some research indicated that there was no association between learner and examiner gender and a minor but highly significant interaction of learner and examiner ethnicity on stations assessing communication skills and ethics (Dewhurst et al., 2007). Additionally, personal and individual preferences and beliefs was found in this research to play a role in how an assessor makes a decision. This sometimes can cause bias in judgement as such feelings and preferences are not related to academic achievements and performances. Assessors in OSCE stations assess candidates face-to-face and therefore can easily recognise candidates ethnicity and cultural background as discussed previously. Assessors need, as highlighted by one of the participants in this research, ‘’*to criticise the behaviour not criticise the person.. You never comment on what you believe the students actions are, you only comment on what you see’’ (assessor no. 18).* Another type of bias was highlighted in this research to happen when assessors who

teach and examine their students recall previous performances or attitudes of candidates. This has the potential to influence the assessor' judgements and cause inconsistency among assessors. In addition to the need to clarify to all assessors what type and level of performance they would expect from candidates, it is ideal to increase their self-awareness about the possible bias when they teach and assess the same candidate.

Categorisation, as in stereotyping, avoids the cognitive resources used to monitor a candidate's category-consistent behaviour and serves as energy-saving or resource-preserving mental devices (Allport, 1954). Stereotypes, therefore, help to simplify perception, judgement, and action. It was found in this research that assessors could face some challenges such as fatigue and boredom while they examine candidates. Assessors, as information processors, when challenged by limitations would necessitate compromises and shortcuts. As mentioned earlier, reluctance and leniency could be influenced by some factors such as fatigue, mood and motivation. Some assessors in this research did declare some level of reluctance and leniency when they are not in a good mood. Similarly, and may be in order to simplify judgement, some assessors in this research did stereotype. Whenever perceivers/assessors are not at their optimal time of day (Bodenhausen, 1990), or are under time pressure (Dijksterhuis & van Knippenberg, 1995; Kruglanski & Freund, 1983), stereotypes were activated. In this research some assessors stereotyped giving general impressions about all male or all female students to possess certain characteristics risking assessment to be biased. For instance, one assessor in this research made a general impression about all female students that they are better than male students in some features such as empathy and eye contact.

In contrast, differences in judgements, as with assessors being idiosyncratic, can sometimes be meaningful. It was found in this research that different assessors made

different interpretations about a single action or scenario. Performance assessments has been perceived as social constructions or interpretations, rather than absolute, objective truths (Johnston, 2004). It is proposed that assessors in work settings develop personal constructs or 'theories' of efficient job performance overall (Ginsburg et al., 2010). Performance theories progress and advance through professional experience, socialisation and training. As a result, the content of performance theories is expected to differ from one assessor to another, causing variant levels of assessor idiosyncrasy (Uggerslev & Sulsky, 2008). Consequently, the notion, underlying psychometric assessment theory, of the reality of a single true score, is challenged by the assessment that is framed in socio-cultural constructivist theories. Experienced assessors in this research were more able to form comprehensive interpretations of performance. In addition, this research found that assessors valued or paid a different degree of attention to different aspects of the two students' performances while observing them. This resulted in different definitions among the participants of this research of what determines quality. This variance in interpretations could be suggested to be used as a valuable feedback to candidates. Since the OSCE can be used to provide feedback to learners in both summative and formative ways (Harden et al., 2015), it would be ideal to use such meaningful and rich feedback to further develop candidates' skills and future learning.

Impact summary

It is the responsibility of medical schools to make assessors aware of what they should expect from candidates. Different expectations were found to be a reason for inconsistency in judgements among assessors. In addition, assessors need to be familiar with the case, task questions, marking sheet, and their role as assessors. This would suggest ideal selection, training and distribution of assessors.

Increasing assessors' self-awareness of different issues that can make their judgements biased is important. For instance, assessors should be aware that inconsistency in patients' performance was found in this research to influence inter-rater reliability. Seeking patient satisfaction does not always provide accurate information about how both real and simulated patients really feel about a candidate.

Inconsistency among assessors can sometimes be meaningful and valid. Delandshere and Petrosky (1994, p. 16) declared: ''Judges' values, experiences, and interests are what makes them capable of interpreting complex performances, but it will never be possible to eliminate those attributes that make them different, even with extensive training''. Therefore, it is suggested that this meaningful variance is used to provide valuable feedback to candidates using the OSCE as both a formative and summative assessment instrument.

C- Patient-related behaviours

The OSCE uses direct observation of candidates communicating and examining real or simulated patients. The third character in the OSCE, patient, was found in this research to influence the other two characters, student and assessor. ''Impressions are subject to variables and contextual factors beyond the candidate himself or herself'' (Gingerich et al., 2011, p. 52). One of these variables was found in this research to be the patient. It was highlighted in this study that it is difficult even for the same simulated patient to give the same performance and act in the same way to each candidate. 'Performance drift' can occur when one case is played by the same simulated patient over a long period of time (McKinley & Boulet, 2004) adversely affecting the process of the assessment (Norcini &

McKinley, 2007). As a result, it is essential to carefully choose simulated patients, train them extensively, and develop an ongoing quality assurance program (Boulet et al., 2002). However, the patients can also be seen as 'fallible' and confront challenges such as fatigue, memory and concentration issues. This can ultimately influence both candidates and assessors and cause inconsistency in judgements among examiners.

The simulated patients can themselves reliably rate candidate's performance with respect to history taking and physical examinations (Epstein, 2007). The medical students in Leeds are assessed by both examiners and patients. Therefore, what is true for assessor might be true for patients. Patients are able to identify candidates' gender and ethnicity as they see each other during an OSCE station. In a similar way to OSCE assessors, some simulated patients can be biased. It was found in this research that some patients can be harsh or very generous in how they deal with a candidate. This can be observed and seen in their satisfaction and in how they cooperate with the candidate in the station. Furthermore, racism was found in this research as a possible reason that could make it difficult for some candidates to perform in an optimal way. All these reasons can have an influence on assessors' judgement and lead to inconsistency among them in making their decisions.

In addition, the difference in culture between the patient and the student can play another role in increasing inconsistency among assessors. It was found in this study that a candidate who comes from a different culture might not be familiar with some culture-related behaviours that a patient might display and expects the student to respond. Such inability to respond to certain behaviours can be interpreted differently and could influence both the assessor and the patient's judgements. The OSCE, as one assessor described, is an *''interplay between patient and doctor. The way some patients behave*

and the way they want response to them is hard to pick up if you are not used to picking them culturally'' (assessor no. 14). Candidates and patients are allowed, by non-verbal communication, to gauge responses, to contextualise the meaning of verbal utterances, and to communicate a "hidden agenda" (Hall et al., 1981; Ishikawa, et al., 2006). However, it was highlighted in this research that simulated patients would be better adapting their behaviours and answers to different students if they are trained well to do so.

Impact summary

Since the patients are asked to assess candidates' performance in the context of this study, it might be possible to say that what is true for assessors is true for patients. Therefore, training of simulated patients should not just focus on them as patients, but also on their responsibility as assessors. They might stereotype or be biased, and increasing their self-awareness about such issues can be helpful. In addition, issues like adaptation and cultural and linguistic differences should be included in training courses to ensure them being consistent with all candidates. The OSCE was introduced to ensure all candidates receive the same test and experience. This same experience can be influenced by the patients being inconsistent due to several factors that need to be taken into consideration.

D- Organisational and environmental factors

This chapter has so far focused on the behaviours of the three main characters in the OSCE – candidates, assessors and patients. However, it has been clearly identified in this research that the context plays a significant role in assessing competence, and the three characters in the OSCE can be influenced by several contextual and organisational

factors. Competence and performance interpretations are to be realised as inherently contextualised (Govaerts & van der Vleuten, 2013). This included, in this study, the preparation of the place of the examination, scenarios, questions and marking sheets. True intra-individual performance disparity could result from changes in the individual (e.g. due to motivation, fatigue, changing levels of competence) as well as changes in the context (Govaerts & van der Vleuten, 2013; Sturman et al., 2005).

The OSCE, as described earlier, is based on the principles of objectivity and simulation which could help enhance the assessment of candidate's performance against predefined standards and criteria and using standardised scoring schemes by trained examiners (Harden et al., 2015). The credibility of such a standard, as it involves judgement, would vary depending on who sets the standards, the characteristics of the methods used, and the outcome (Norcini, 2003). Some of the assessors in this research highlighted the necessity of using well written and clear scoring schemes to help them assess reliably. It has been shown in this research that a lack of anchors for assessors can actually be quite tough, and that too open a scoring format cause difficulties. The word 'adequate', for example, might be interpreted differently by different assessors either because of the different expectations assessors have or because of individual differences between assessors in interpretation. The literature suggests that making detailed checklists might not always help in improving objectivity as the possibility of cognitive load increases (Tavares & Eva, 2013). However, some assessors in this research highlighted that if the marking sheet is quite broad, the assessor is left not quite sure which way they should mark. Therefore, some balance needs to be taken into consideration. It is also possible to allow assessors to write down a few sentences before they make their general judgement, as discussed earlier, to help them articulate what is not listed in the marking sheet. In

addition, it is worth reiterating that all assessors need to be briefed about what they should expect from candidates to eliminate differences in expectations.

An appropriate assessment length has been recognised to greatly increase assessment reliability (Swanson et al., 1995). The use of multiple examiners across different cases with sufficient testing time also has the potential to achieve adequate increase reliability (Norcini et al., 1985; Swanson, 1987). It was found in this research that the time dedicated to each station was considered by some assessors as short. Examiners in the OSCE are expected to experience mental workload that is higher than that which occurs in other routine clinical work (Byrne et al., 2014). This was confirmed by some assessors in this research due to the need to assess candidates' skills and knowledge in a short period of time. This shortness in time can cause some difficulty for both the student and the assessor and place some pressure on them. This pressure placed on assessors was seen in this study as a challenge that could increase the pressure of self-discipline on assessors. Moreover, a statement like *''one weakness of the OSCE is that students will be doing so many things in a very short period of time'' (assessor no. 2)* can suggest that some assessors might tend to be more lenient with candidates because of this shortness in time. In addition, such shortness in time, seen by the assessors in this study, could increase subjectivity while assessing candidates. A statement like *''sometimes I do have the feeling that OSCE is not as objective as it could be depending on the station and also depending on the timing of the OSCE'' (assessor no. 15)* might give an explanation why some judgements tend to be more subjective. It was discussed earlier that categorisation saves a lot of mental efforts when perceivers face mental challenges. Assessors in challenging situations such as assessing several skills in a short period of time,

accompanied by background noises or high room temperature, might activate categorisation or bias.

Impact summary

The importance of the context of the exam must not be underestimated as the three characters in the OSCE were found in this research to be influenced by it. When everything is in place it impresses everyone in the station, and it helps increase their focus on their tasks. A statement like *''for me it is often the external things which I find most frustrating rather than necessarily the internal things'' (assessor no. 5)* shows how important the preparation of the context on the assessment process can be. Background noises or high room temperature are two examples that were found in this study to decrease concentration. In addition, breaks between stations was seen as essential for assessors to maintain focus and concentration. Assessors might get bored when they listen to the same thing all day, and so it is better to switch between them to observe and listen to something new. Well written marking sheets could also help assessors avoid some negative influence caused by tiredness or boredom. The time allocated for each station was considered short by some assessors. A statement like *''It is fairly short time really. You know in a hospital setting you will not necessary be limited to that amount of time'' (assessor no. 3)* can suggest that candidates are asked to do many things in a short period of time. It is suggested that a slight increase in time or decrease in the number of questions asked would help alleviate the negative impacts caused by such a lack of time. However, the practicability and manageability of this step needs to be considered and studied.

E- Implications

It does not need to be voiced to be counted.

The non-verbal behaviours of the three 'characters' in the OSCE (student, patient and assessor), and the environment in which it is situated, make significant contributions to global ratings and contribute to the multiple factors that reduce inter-rater reliability. This has importance in station and scoring format design, assessor and patient selection and training and the ongoing research into the assessor decision-making in high stakes performance tests.

Candidates

It is important to note that competence cannot be restricted by a list of skills. Some assessors in this research, for example, paid more attention to examining how candidates adapt themselves to different scenarios and situations, and not just on perfection. Such a skill will always depend on the context, task presented and patient. Therefore, such a skill changes from one context to another. It was discussed earlier in the literature that meaningful assessment should allow candidates to show skills that are beyond what is listed in the marking scheme. Therefore, assessment needs to consider that candidates' skills cannot always be counted or listed prior to each exam. Albert Einstein (as cited in McFarlane, 2004) once said, ''Everything that can be counted does not necessarily count; everything that counts cannot necessarily be counted.'' Different examples, found in this study, show that assessors' judgements were influenced by different skills and non-verbal behaviours that sometimes are not listed in the marking sheet. It is also suggested that the non-verbal behaviours found in this research are highlighted to students, in consultation skills courses, as such behaviours played a role in forming assessors' judgements.

It is also found in this research that the OSCE needs to be introduced to candidates before they experience it in a summative way. The concept of role play was found in this study to be an issue that might cause difficulty to some students. It would be ideal first to introduce the OSCE as a formative assessment instrument. This will ensure that students receive feedback and get to know the OSCE in a practical way. Students can ask questions and seek help whenever needed regarding the assessment process and the usual procedures taking place in summative examinations.

Assessors

Assessors were found in this study to have different expectations about the level of competence candidates need to show. Such differences in expectations needs to be taken into consideration as a potential for inconsistency in judgements. It is important that all assessors possess similar expectations about the general performance of candidates. In addition, it is important that all assessors are familiar with their roles as assessors, the case and task presented, and with the marking sheet. Furthermore, assessors need to have self-awareness about issues found in this research that could make judgements biased such as comparisons, categorisation or stereotype. Although it might not be possible to completely eliminate such effects, possessing self-awareness about these issues can help alleviate their influence.

Using the patient as a frame of standard was also found in this research as a possible potential for inconsistency in judgement among assessors. Patients were found in this research to sometimes be inconsistent due to several reasons such as fatigue or boredom. Building judgements on inconsistent behaviours will ultimately produce inconsistent

judgements. Some assessors might not be aware of this point, and therefore it is worth mentioning in training courses.

The idiosyncrasy of assessors can produce rich and meaningful feedback in formative and summative assessments. It is very possible, and highly encouraged, based on this study, to use the OSCE as a formative assessment tool in medical education to provide immediate feedback and to further build constructive alignment. It is also essential to distinguish between inaccuracy and idiosyncrasy when it comes to assessors' judgements. Variances in an assessor's interpretation and scoring of performance could be equally valid (Landy & Farr, 1980) and meaningful (Lance et al., 2008). Delandshere and Petrosky (1994, p.16) declared: ''judges' values, experiences, and interests are what makes them capable of interpreting complex performances, but it will never be possible to eliminate those attributes that make them different, even with extensive training''. Information Integration refers to assessors explaining the valence of their comments in their own unique narrative terms, usually leading to global impressions formation. While assessors make judgements, they explain and probably mentally represent the valence of those judgements in unique narrative terms. These unique narrative and global judgements, in turn, are converted into the assessment scale to produce scores for each individual domain (Yeates et al., 2013b), which ultimately lead to inconsistency and low inter-rater reliability. This research, therefore, would support allowing the examiners to narrate their judgements and feedback, instead of only ticking boxes and giving global judgements. A recent study by Harrison et al. (2015) reported that audio feedback after summative OSCEs was seen as meaningful by candidates. Such feedback and narration can be a few sentences that could provide meaningful feedback and explanations of decisions. This type of feedback can be delivered in both summative and formative assessments.

Although its use is established in summative assessments, the OSCE can also be adapted for formative assessments.

Patients

It would be possible to say that what is true for assessors can be true for patients. Training of simulated patients should also include elements regarding the process of assessments they conduct during OSCE stations. Such elements are related to what the patients are usually asked to assess and how to balance between being a patient and an assessor at the same time. Increasing their self-awareness about the issues that can affect their consistency is important to alleviate any negative impact these issues possess on assessment and candidates' performance.

Institution

In addition to taking into consideration the previous mentioned point, the context of the exam and how it is prepared can play an important role in the assessment process as it is found in this study to influence all the three characters in the OSCE. Ensuring an ideal environment for assessment helps everyone to concentrate on their tasks. Breaks between stations were found very important to maintain concentration. Furthermore, assessors switching between stations was highlighted as necessary to avoid listening to the same topic and questions over and over which can cause some kind of boredom, and therefore decrease assessors' concentration. However, this needs careful planning to ensure assessors being familiar with the task, questions and marking sheet as discussed earlier. Furthermore, it is suggested that the time allocated to each station be slightly increased or the questions asked in each station decreased if doable and applicable. Finally, the marking sheet itself can be a distracter if it is not clear and well written. After ticking

boxes, and before or after making a global marking, assessors could also narrate their feedback and explanations of their judgements to avoid missing rich and valuable feedback.

Finally, it is important to note that the OSCE is still seen as a powerful assessment instrument. The previous discussed implications could help increase the output of this assessment method and better understand some issues related to inter-rater reliability. Big national exams that seek to standardise and adjust for every little bit of variance to avoid candidate appeals risk losing so much of that unique, rich and diverse feedback provided by assessors.

Chapter 6 Conclusion

'It does not need to be voiced to be counted.'

Whilst OSCEs are a well-recognised format for assessing clinical competence, an increasing body of research focuses on the factors that contribute to differences in assessors' judgement in performance assessment. Perspectives from social and psychosocial research have explored factors influencing these differences, but less attention has been paid to non-verbal behaviours of candidates, assessors and patients that could influence assessors' judgements during OSCEs. This research investigated how non-verbal behaviour could influence assessors' global marking when examining undergraduate medical students using OSCEs.

In the 'theatre of performance' of the OSCE, all the characters contribute to variance – and thus (unlike many other studies) this research does not just focus on one character or another, but all and the environment. The nonverbal behaviours of the three 'characters' in the OSCE (student, patient and assessor), and the environment in which it is situated, make significant contributions to global ratings and contribute to the multiple factors that influence inter-rater reliability. This is important in station and scoring format design, assessor selection and training and the ongoing research into assessor decision-making in high stakes performance tests.

Finally, it is worth mentioning that whilst the OSCE has been utilized as a summative assessment instrument, it has been far less well used for formative purposes. This is a missed opportunity as possibly the very rich data and narrative that come from it make it ideal for use in formative assessments.

Study limitations

This study used videos, instead of real face-to-face assessments, to collect the required data. Although it is recognised as efficient in obtaining the required data, it might not be as efficient as real face-to-face assessments. The videos were quite short, around 2 mins, which do not represent real OSCE stations. Therefore, the assessors in this research might not have had enough time to observe what usually takes place in a real OSCE station. In addition, the videos showed one angle of the room showing only the candidate's face and body. Finally, all the information gathered about the three characters and environment, in this research, was based on the views of one character, the assessor. Different perspectives obtained from the other two characters could add valuable information, especially the views from patients as assessors.

Future research

It is suggested that attention is paid to the other two characters regarding the process of assessment in the OSCE and the effects of non-verbal behaviours on performance and assessment. The patient as assessor and how they balance between being patients and assessors at the same time is worth investigating.

References

Abele, A. E., Cuddy, A. J. C., Judd, C. M., & Yzerbyt, V. Y. (2008). Fundamental dimensions of social judgment. *European Journal of Social Psychology, 38*, 1063-1065

Abele, A. E., & Wojciszke, B. (2007). Agency and Communion From the Perspective of Self Versus Others. *Journal of Personality and Social Psychology, 93*(5), 751-763

Adibi, H. (2014). Social Perspectives on Health: mHealth Implications. In S. Adibi (Ed.), *MHealth Multidisciplinary Verticals* (pp. 219-236). Washington: Taylor & Francis Group.

Albanese, M. (2000). Challenges in using rater judgements in medical education. *Journal of Evaluation in Clinical Practice, 6*(3), 305-319.

Albanese, M. (1999). Rating education quality: Factors in the erosion of professional standards. *Academic medicine, 74*, 652-658.

Albright, L., Malloy, T. E., Dong, Q., Kenny, D. A., Fang, X., Winquist, L., & Yu, D. (1997). Cross-Cultural Consensus in Personality Judgments. *Journal of Personality and Social Psychology, 72*(3), 558-569.

Allen, D., Benner, P., & Diekelmann, N. (1986). Three paradigms for nursing research: methodological implications. In P. Chinn (Ed.), *Nursing Research Methodology: Issues and Implementation* (pp. 23-38). Rockville, MD: Aspen.

Allport, G. W. (1954). *The nature of prejudice*. Boston, Mass, Cambridge: Addison-Wesley Pub. Co.

Alves de Lima, A., Barrero, C., Baratta, S., Castillo Costa, Y., Bortman, G., Carabajales, J., & al., e. (2007). Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Excercise (Mini-CEX for cardiology residency training). *Medical Teacher, 29*(8), 785-790.

Alves de Lima, A., Conde, D., Costabel, J., Corso, J., & Van der Vleuten, C. (2011). A laboratory study on the reliability estimations of the mini-CEX. *Advances in Health Sciences Education, 18*(1), 5–13

Alves de Lima, A., Henquin, R., Thierer, J., Paulin, J., Lamari, S., Belcastro, F., & van der Vleuten, C. (2005). A qualitative study of the impact on learning of the mini clinical evaluation exercise in postgraduate training. *Medical Teacher, 27*(1), 46-52.

Amin, Z., & Khoo, H. (2003). Overview of Assesment and Evaluation *Basics in Medical Education* (pp. 251-260). Singapore: World Scientific Publishing Company.

Anastasi, A. (1988). *Psychological Testing*. New York: MacMillan Publishing Company.

Anderson, N. (1968). A simple model for information intgration In R. Abelson, E. Aronson, W. McGuire, T. Newcomb, M. Rosenberg & P. Tannenbaum (Eds.), *Theories of cognitive consistency: A sourcebook* (pp. 731-743). Chicago: Rand McNally.

Andersen, S., & Cole, S. (1990). ''Do I know you?'': The role of significant others in general social perception. *Journal of Personality and Social Psychology, 59*, 384-399.

Andersen, S. M., Klatzky, R. L., & Murray, J. (1990). Traits and Social Stereotypes: Efficiency Differences in Social Information Processing. *Journal of Personality and Social Psychology, 59*(2), 192-201.

Anderson, L. W. & Krathwhol, D. R. (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Blooms' Taxonomy of Educational Objectives*. New York: Addison Wesley Longman.

Archer, J., McGraw, M., & Davies, H. (2010). Assuring validity of multisource feedback in a national programme. *Archives of disease in childhood, 95*(5), 330-335.

Asch, S. E. (1946). Forming impressions of personality. *The Journal of Abnormal and Social Psychology, 41*(3), 258-290.

Asch, S. E., & Zukier, H. (1984). Thinking about persons. *Journal of Personality and Social Psychology, 46*(6), 1230-1240.

Baddeley, A. (1994). The Magical Number Seven: Still Magic After All These Years? *Psychological Review, 101*(2), 353-356.

Baig, L. A., Violato, C., & Crutcher, R. A. (2009). Assessing clinical communication skills in physicians: are the skills context specific or generalizable. *BMC medical education, 9*(1), 22-22.

Barbour, R. S. (2001). Checklists For Improving Rigour In Qualitative Research: A Case Of The Tail Wagging The Dog? *BMJ: British Medical Journal, 322*(7294), 1115-1117.
Bargh, J. (1989). Conditional automaticity: Varieties of automatic influence in social perception and cognition. In J. Uleman & J. Bargh (Eds.), *Unintended thought* (pp. 3-51). New York: Guilford Press.
Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54*(7), 462-479.
Bargh, J. A., & Ferguson, M. J. (2000). Beyond behaviorism: On the automaticity of higher mental processes. *Psychological Bulletin, 126*(6), 925-945
Batalden, P., Leach, D., Swing, S., Dreyfus, H., & Dreyfus, S. (2002). General Competencies And Accreditation In Graduate Medical Education. *Health Affairs, 21*(5), 103-111
Beal, D. J., Weiss, H. M., Barros, E., & MacDermid, S. M. (2005). An episodic process model of affective influences on performance. *Journal of Applied Psychology, 90*, 1054-1068
Beauvois, J.-L. o., & Dubois, N. (2009). Lay Psychology and the Social Value of Persons. *Social and Personality Psychology Compass, 3*(6), 1082-1095
Becker, G. A., & Miller, C. E. (2002). Examining Contrast Effects in Performance Appraisals: Using Appropriate Controls and Assessing Accuracy. *The Journal of Psychology, 136*(6), 667-683.
Becker, G. A., & Villanova, P. (1995). Effects of Rating Procedure and Temporal Delay on the Magnitude of Contrast Effects in Performance Ratings. *The Journal of Psychology, 129*(2), 157-166.
Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. T. (2013). Expertise in performance assessment: assessors' perspectives. *Advances in Health Sciences Education, 18*(4), 559-571.
Berwick, D. M. (2008). The Science of Improvement. *JAMA, 299*(10), 1182-1184.
Biggs, J. & Tang, C. (2007). *Teaching for Quality Learning at University* (3rd ed.). Maidenhead: Open University Press.
Black, P. & Wiliam, D. (1998). Inside the Black Box: Raising Standards Through Classroom Assessment. *Phi Delta Kappan, 80*(2), 139-144.
Blair, I. V., & Banaji, M. R. (1996). Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology, 70*, 1142–1163.
Bleakley, A. (2010). Blunting Occam's razor: aligning medical education with studies of complexity: Medical education and complexity. *Journal of Evaluation in Clinical Practice, 16*(4), 849-855
Bless, H., Fiedler, K., & Strack, F. (2004). *Social Cognition: How Individuals Construct Social Reality*. New York, NY: Psychology Press.
Bodenhausen, G. V. (1988). Stereotypic Biases in Social Decision Making and Memory: Testing Process Models of Stereotype Use. *Journal of Personality and Social Psychology, 55*(5), 726-737.
Bodenhausen, G. V. (1990). Stereotypes as Judgmental Heuristics: Evidence of Circadian Variations in Discrimination. *Psychological Science, 1*(5), 319-322.
Bodenhausen, G. V. (1993). Emotion, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping In D. Mackie & D. Hamilton (Eds.), *Affect, cognition, and stereotyping: Interactive processes in group perception* (pp. 13-37). San Diego, CA: Academic Press.
Bodenhausen, G. V., Kramer, G. P., & Süsser, K. (1994). Happiness and Stereotypic Thinking in Social Judgment. *Journal of Personality and Social Psychology, 66*(4), 621-632.
Bodenhausen, G. V., & Lichtenstein, M. (1987). Social Stereotypes and Information-Processing Strategies: The Impact of Task Complexity. *Journal of Personality and Social Psychology, 52*(5), 871-880.
Bodenhausen, G. V., Sheppard, L. A., & Kramer, G. P. (1994). Negative affect and social judgment: The differential impact of anger and sadness. *European Journal of Social Psychology, 24*(1), 45-62.

Bodenhausen, G. V., & Wyer, R. S. (1985). Effects of Stereotypes on Decision Making and Information-Processing Strategies. *Journal of Personality and Social Psychology, 48*(2), 267-282.

Bogdan, R. C., & Biklen, S. K. (2006). *Qualitative research for education: An introductory to theory and methods* (5th ed.). Needham Heights, MA: Allyn and Bacon.

Bordage, G. (2009). Conceptual frameworks to illuminate and magnify. *Medical education, 43*(4), 312-319.

Borman, W. C. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an army officer sample. *Organizational Behavior and Human Decision Processes, 40*(3), 307-322.

Borrell-Carrió, F., & Epstein, R. M. (2004). Preventing errors in clinical practice: a call for self-awareness. *Annals of family medicine, 2*(4), 310-316

Boud, D. (1995). *Enhancing learning through self assessment*. London: Kogan Page.

Boud, D. (2000). Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151-167

Boulet, J. (2005). Generalizability theory: Basis. In S. Everitt & C. Howell (Eds.), *Encyclopedia of Statistics in Behavioural Science* (2nd ed., pp. 704-711). Chichester: John Wiley & Sons.

Boulet, J. R., McKinley, D., Norcini, J., & Whelan, G. (2002). Assessing the comparability of standardized patient and physician evaluations of clinical skills *Advances in Health Sciences Education: Theory and Practice, 7*(2), 85-97.

Boulet, J. R., McKinley, D. W., Whelan, G. P., & Hambleton, R. K. (2003). Quality Assurance Methods for Performance-Based Assessments. *Advances in Health Sciences Education, 8*(1), 27-47

Boulet, J. R., Smee, S., Dillon, G., & Gimble, J. (2009). The use of standardized patient assessments for certification and licensure decisions *Simulation in Healthcare, 4*, 35-42

Bourne, E. (1977). Can we describe an individual's personality? Agreement on stereotype versus individual attributes. *Journal of Personality and Social Psychology, 35*(12), 863-872.

Boursicot, K., Etheridge, L., Setna, Z., Sturrock, A., Ker, J., Smee, S., & Sambandam, S. (2011). Performance in assessment: Consensus statement and recommendations from the Ottawa conference. *Medical Teacher, 33*, 370–383

Boursicot, K., Roberts, T., & Burdick, W. (2014). Structured assessments of clinical competence. In T. Swanwick (Ed.), *Understanding Medical Education Evidence, Theory and Practice* (pp. 246-258). Chichester, UK: John Wiley & Sons.

Braddock Iii, C. H., Fihn, S. D., Levinson, W., Jonsen, A. R., & Pearlman, R. A. (1997). How Doctors and Patients Discuss Routine Clinical Decisions: Informed Decision Making in the Outpatient Setting. *Journal of General Internal Medicine, 12*(6), 339-345.

Branch, W. T. & Paranjape, A. (2002). Feedback and reflection: teaching methods for clinical settings. *Academic Medicine, 77*, 1185-1188.

Brannick, M. T., Erol-Korkmaz, H. T., & Prewett, M. (2011). A systematic review of the reliability of objective structured clinical examination scores. *Medical education, 45*(12), 1181-1189

Brazeau, C., Boyd, L., & Crosson, J. (2002). Changing an existing OSCE to a teaching tool: the making of a teaching OSCE. *Academic medicine : journal of the Association of American Medical Colleges, 77*(9), 932

Brewer, M. (1988). A dual process model of impression formation. In T. Srull & R. Wyer (Eds.), *Advances in social cognition* (pp. 1-36). Hillsdale, NJ: Erlbaum.

Brewer, M., & Harasty Feinstein, A. (1999). Dual processes in the cognitive representation of persons and social categories. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology*. New York: Guilford Press.

Brewer, M. B., Dull, V., & Lui, L. (1981). Perceptions of the elderly: Stereotypes as prototypes. *Journal of Personality and Social Psychology, 41*(4), 656-670.

Brewer, N. T., & Chapman, G. B. (2003). Contrast Effects in Judgments of Health Hazards. *The Journal of Social Psychology, 143*(3), 341-354.

British Medical Association. (2006) *Examining equality: a survey of royal college examinations*. London: BMA.
Britten, N. (1995). Qualitative interviews in medical research. *British Medical Journal, 311*, 251-253.
Britten, N. (2005). Making sense of qualitative research: a new series. *Medical education, 39*(1), 5-6
Brutus, S. (2010). Words versus numbers: A theoritical exploration of giving and receiving narrative comments in performance appraisal. *Human Resource Management Review, 20*(2), 144-157.
Bulawa, P. (2014). Adapting grounded theory in qualitative research: reflections from personal experience. *International Research in Education, 2*, 79-114
Bunniss, S., & Kelly, D. R. (2010). Research paradigms in medical education research. *Medical education, 44*(4), 358-366.
Burt, C. (1936). The analysis of examination marks. In P. Hartog & E. Rhodes (Eds.), *The marks of examiners* (pp. 309-310). London: MacMillan.
Byrne, A., Tweed, N., & Halligan, C. (2014). A pilot study of the mental workload of objective structured clinical examination examiners. *Medical education, 48*(3), 262-267.
Cacamese, S. M., Elnicki, M. & Speer, A. J. (2007). Grade inflation and the internal medicine subinternship: a national survey of clerkship directors. *Teaching & Learning in Medicine, 19*(4), 343-346.
Calsyn, R. J. (2000). A checklist for critiquing treatment fidelity studies. *Mental Health Services Research, 2*(107-113).
Cantor, N., & Mischel, W. (1977). Traits as prototypes: Effects on recognition memory. *Journal of Personality and Social Psychology, 35*(1), 38-48.
Cantor, N., & Mischel, W. (1979). Prototypes in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 3-52). New York: Academic Press.
Cardy, R. L., Bernadin, H. J., Abbott, J. G., Senderak, M. P., & Taylor, K. (1987). The effects of individual performance schemata and dimension familiarization on rating accuracy. *Journal of Occupational Psychology, 60*(3), 197.
Caris-Verhallen, W., Kerkstra, A. & Bensing, J. (1999). Non-verbal behavior in nurse-elderly patient communication. *Journal of Advanced Nursing, 29*, 808-818.
Carlston, D., & Smith, E. (1996). Principles of mental representation In E. Higgins & W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* New York: The Guilford Press.
Carr, Woodson, Woolf, K., Woolf, J., Cave, J., Greenhalgh, T., & Dacre, J. (2008). Ethnic Stereotypes and the Underachievement of UK Medical Students from Ethnic Minorities: Qualitative Study. *BMJ: British Medical Journal, 337*(7670), 611-615.
Chapman, G. B., & Bornstein, B. H. (1996). The More You Ask For, the More You Get: Anchoring in Personal Injury Verdicts. *Applied Cognitive Psychology, 10*(6), 519-540.
Charmaz, K. (2014). *Constructing grounded theory* (Vol. 2nd). London: SAGE.
Charmaz, K. (1995). Grounded theory. In J. Smith, R. Harre & L. Langenhove (Eds.), *Rethinking Methods in Psychology* (pp. 27-49). London: Sage.
Charmaz, K. (1983). The grounded theory method: An explication and interpretation. In R. M. Emerson (Ed.), *Contemporary field Research* (pp. 109-126). Boston: Little, Brown and Company.
Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152.
Chiu, C.-y., Hong, Y.-y., Lam, I. C.-m., Fu, J. H.-Y., Tong, J. Y.-y., & Lee, V. S.-l. (1998). Stereotyping and Self-Presentation: Effects of Gender Stereotype Activation. *Group Processes & Intergroup Relations, 1*(1), 81-96.
Clauser, B. E., Clyman, S. G. & Swanson, D. B. (1999). Components of rater error in a complex performance assessment. *Journal of Educational Measurement, 36*(1), 29-45.
Clauser, B. E., Harik, P., & Clyman, S. G. (2000). The Generalizability of Scores for a Performance Assessment Scored with a Computer-Automated Scoring System. *Journal of Educational Measurement, 37*(3), 245-262.

Clauser, B. E., Harik, P., Margolis, M. J., Mee, J., Swygert, K., & Rebbecchi, T. (2008). The generalizability of documentation scores from the USMLE Step 2 Clinical Skills examination. *Academic medicine, 83*(10 Suppl), S41.

Clauser, B. E., Margolis, M. J., & Swanson, D. B. (2002). An Examination of the Contribution of Computer-based Case Simulations to the USMLE Step 3 Examination. *Academic medicine, 77*(10 Suppl), S80-S82

Clauser, B. E., Subhiyah, R. G., Nungester, R. J., Ripkey, D. R., Clyman, S. R. & McKinley, D. (1995). Scoring a performance-based assessment by modeling the judgments of experts. *Journal of Educational Measurement, 32*(4), 397-415.

Clauser, B. E., Swanson, D. B. & Clyman, S. G. (1996). The generalizability of scores from a performance assessment of physicians' patient management skills. *Academic Medicine, 71*(10 Suppl), S109-111.

Cleland, J. A., Knight, L. V., Rees, C. E., Tracey, S., & Bond, C. M. (2008). Is it me or is it them? Factors that influence the passing of underperforming students. *Medical education, 42*(8), 800-809.

Clyman, S., Melnick, D., & Clauser, B. (1999). Computer-based case simulations from medicine: Assessing skills in patient management. In A. Tekian, C. McGuire & W. McGaghie (Eds.), *Innovative simulations for assessing professional competence* (pp. 29–41). Chicago: University of Illinois, Department of Medical Education.

Collins, L. G., Schrimmer, A., Diamond, J. & Burke, J. (2011). Evaluating verbal and non-verbal communication skills, in an ethnogeriatric OSCE. *Patient Education and Counseling, 83*(2), 158-162.

Colliver, J. A. (2002). Educational theory and medical education practice: a cautionary note for medical school faculty. *Academic medicine : journal of the Association of American Medical Colleges, 77*(12 Pt 1), 1217-1220.

Colliver, J. A. & Swartz, M. H. (1997). Assessing clinical performance with standardized patients. *JAMA, 278*(9), 790-791.

Cook, D. A., & Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine, 119*, e7-e16

Cook, D. A., & Beckman, T. J. (2009). Does scale length matter? A comparison of nine- versus five-point rating scales for the mini-CEX. *Advances in health sciences education : theory and practice, 14*(5), 655.

Cook, D. A., Beckman, T. J., Mandrekar, J. N., & 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.

Cook, D. A., Dupras, D. M., Beckman, T. J., Thomas, K. G., & Pankratz, V. S. (2009). Effect of Rater Training on Reliability and Accuracy of Mini-CEX Scores: A Randomized, Controlled Trial. *Journal of General Internal Medicine, 24*(1), 74-79.

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13*(1), 3-21.

Costrich, N., Feinstein, J., Kidder, L., Marecek, J., & Pascale, L. (1975). When stereotypes hurt: Three studies of penalties for sex-role reversals. *Journal of Experimental Social Psychology, 11*(6), 520-530.

Cox, K. (1990). No Oscar for OSCE. *Medical education, 24*(6), 540-545

Crabtree, B., & Miller, W. (1999). *Doing Qualitative Research* (2nd ed.). London: Sage.

Crawford, M. T., & Skowronski, J. J. (1998). When Motivated Thought Leads to Heightened Bias: High Need for Cognition Can Enhance the Impact of Stereotypes on Memory. *Personality and Social Psychology Bulletin, 24*(10), 1075-1088.

Criley, J. M., Vukanovic-Criley, J. M., Nelson, W. P., Guevara-Matheus, L., Warde, C. M., Criley, S., . . . Churchill, W. H. (2006). Competency in Cardiac Examination Skills in Medical Students, Trainees, Physicians, and Faculty: A Multicenter Study. *Archives of Internal Medicine, 166*(6), 610-616.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302

Crossingham, G. V., Sice, P. J. A., Roberts, M. J., Lam, W. H., & Gale, T. C. E. (2012). Development of workplace-based assessments of non-technical skills in anaesthesia. *Anaesthesia, 67*(2), 158-164.

Crossley, J., Humphris, G. & Jolly, B. (2002). Assessing health professionals. *Medical Education, 36*(9), 800-804.

Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Medical education, 46*(1), 28-37.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2007). The BIAS Map: Behaviors from intergroup affect and stereotypes. *Journal of Personality and Social Psychology, 92*, 631–648

Cunnington, J. P., Neville, A. J., & Norman, G. R. (1997). The risks of thoroughness: Reliability and validity of global ratings and checklists in an OSCE. In A. J. Scherpbier, C. P. van der Vleuten & J. J. Rethans (Eds.), *Advances in Medical Education* (pp. 143-145). The Netherlands: Kluwer.

Damisch, L., Mussweiler, T., & Plessner, H. (2006). Olympic Medals as Fruits of Comparison? Assimilation and Contrast in Sequential Performance Judgments. *Journal of Experimental Psychology: Applied, 12*(3), 166-178.

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology, 44*(1), 20-33.

Dauphinee, D. (1994). Determining the content of cerification examinations. In D. Newble, B. Jolly & R. Wakeford (Eds.), *The certification and recertification of doctors: issues in the assessment of clinical competence*. Cambridge: Cambridge University Press.

Deadrick, D. L., Bennett, N., & Russell, C. J. (1997). Using hierarchical linear modeling to examine dynamic performance criteria over time. *Journal of Management, 23*(6), 745-757

Deaux, K., & Taynor, J. (1973). Evaluation of male and female ability: Bias works two ways. *Psychological Reports, 32*, 261-262.

Delandshere, G. (2002). Assessment as inquiry. *Teachers College Record, 104*(7), 1461-1484

Delandshere, G., & Petrosky, A. R. (1994). Capturing Teachers' Knowledge: Performance Assessment a) And Post-Structuralist Epistemology, b) From a Post-Structuralist Perspective, c) And Post-Structuralism, d) None of the above. *Educational Researcher, 23*(5), 11-18.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of Complex Performances: Limitations of Key Measurement Assumptions. *Educational Researcher, 27*(2), 14-24.

DeNisi, A. S. (1984). A cognitive view of the performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*(3), 360-396.

DeNisi, A. S. (1996). *A cognitive approach to performance appraisal: a program of research*. New York;London;: Routledge.

Denney, M. L., Freeman, A., & Wakeford, R. (2013). MRCGP CSA: Are the examiners biased, favouring their own by sex, ethnicity, and degree source? *British Journal of General Practice, 63*(616), e718-725

Dent, J. A., & Harden, R. M. (2013). *A practical guide for medical teachers* (Vol. Fourthition.). Edinburgh: Churchill Livingstone.

Denzin, N. (1970). *The research act in sociology: A theoritical introduction to sociological methods*. London: Butterworth.

Denzin, N. K., & Lincoln, Y. S. (2000). *Handbook of qualitative research* (Vol. 2nd). Thousand Oaks, Calif;London;: Sage.

Denzin, N. K., & Lincoln, Y. S. (2005). *The SAGE handbook of qualitative research* (3rd ed.). Thousand Oaks: Sage Publications.

Devine, P. G. (1989). Stereotypes and Prejudice: Their Automatic and Controlled Components. *Journal of Personality and Social Psychology, 56*(1), 5-18.

Devine, P. G., & Baker, S. M. (1991). Measurement of Racial Stereotype Subtyping. *Personality and Social Psychology Bulletin, 17*(1), 44-50.

Dewhurst, N. G., McManus, C., Mollon, J., Dacre, J. E., & Vale, A. J. (2007). Performance in the MRCP(UK) Examination 2003-4: analysis of pass rates of UK graduates in relation to self-declared ethnicity and gender. *BMC medicine, 5*(1), 8-8.

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical education, 40*(4), 314-321.

Dijksterhuis, A., & Knippenberg, A. v. (1995). Memory for Stereotype-Consistent and Stereotype-Inconsistent Information as a Function of Processing Pace. *European Journal of Social Psychology, 25*(6), 689-693.

DiMatteo, M. (2004). The role of effective communication with children and their families in fostering adherence to pediatric regimens. *Patient Education & Counseling, 55*, 339-344.

DiMatteo, M. R., Taranta, A., Friedman, H. S. & Prince, L. M. (1980). Predicting patient satisfaction from physicians' nonverbal communication skills. *Medical Care, 18*(4), 376-387.

Dipboye, R. L., Fromkin, H. L., & Wiback, K. (1975). Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. *Journal of Applied Psychology, 60*(1), 39-43.

Donato, A. A., Pangaro, L., Smith, C., Rencic, J., Diaz, Y., Mensinger, J., & Holmboe, E. (2008). Evaluation of a novel assessment form for observing medical residents: a randomised, controlled trial. *Medical education, 42*(12), 1234-1242.

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical education, 38*(9), 1006-1012.

Downing, S. M. (2005). Threats to the validity of clinical teaching assessments: What about rater error? *Medical education, 39*(4), 353-355.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical education, 37*(9), 830-837

Downing, S. M. & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327-333.

Dreyfus, H. L. (2001). *On the internet*. London: Routledge.

Dudek, N. L., Marks, M. B., & Regehr, G. (2005). Failure to Fail: The Perspectives of Clinical Supervisors. *Academic medicine, 80*(Supplement), S84-S87.

Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of Blacks. *Journal of Personality and Social Psychology, 34*(4), 590-598.

Dunning, D. & Lange, B. (1987). The effect of feedback on student use of interpersonal communication skills. *Journal of Dental Education, 51*, 594-596.

Durning, S., & Artino, A. (2011). Situativity theory: a perspective on how participants and the environment can interact: AMEE Guide no. 52. *Medical Teacher, 33*(3), 188-199.

Durning, S. J., Artino, J. A. R., Pangaro, L. N., van der Vleuten, C., & Schuwirth, L. (2010). Perspective: redefining context in the clinical encounter: implications for research and training in medical education. *Academic medicine : journal of the Association of American Medical Colleges, 85*(5), 894-901.

Durning, S. J., Cation, L. J., Markert, R. J., & Pangaro, L. N. (2002). Assessing the reliability and validity of the mini-clinical evaluation exercise for internal medicine residency training. *Academic medicine : journal of the Association of American Medical Colleges, 77*(9), 900-904.

Durning, S. J., Vleuten, C. P. M. v. d., Schuwirth, L., & Artino, A. R. (2013). Clarifying assumptions to enhance our understanding and assessment of clinical reasoning. *Academic medicine, 88*(4), 442-448

Ebel, R. L. & Frisbie, D. A. (1986). *Essentials of educational measurement*. Englewood Cliffs, New Jersey: Prentice-Hall.

Elliot, D. L., & Hickam, D. H. (1987). Evaluation of Physical Examination Skills: Reliability of Faculty Observers and Patient Instructors. *JAMA, 258*(23), 3405-3408.

Elliot, D. L. & Hickam, D. H. (1987). Evaluation of physical examination skills. Reliability of faculty observers and patient instructors. *JAMA, 258*(23), 3405-3408.
Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: an analysis of clinical reasoning*. Cambridge, MA: Harvard University Press.
Elton, L. (1987). *Teaching in Higher Education: Appraisal and Training*. London: Kogan Page.
Elton, L. (2005). Designing assessment for creativity: an imaginative curriculum guide Retrieved March 2, 2012, from http://www.heacademy.ac.uk/2841.htm
Engeström, Y., & Sannino, A. (2010). Studies of expansive learning: Foundations, findings and future challenges. *Educational Research Review, 5*(1), 1-24.
Epley, N., & Gilovich, T. (2001). Putting Adjustment Back in the Anchoring and Adjustment Heuristic: Differential Processing of Self-Generated and Experimenter-Provided Anchors. *Psychological Science, 12*(5), 391-396.
Epstein, R. M. (2007). Assessment in medical education. *New England Journal of Medicine, 356*(4), 387-396.
Epstein, R. M. (1999). Mindful practice. *The Journal of the American Medical Association (JAMA), 282*, 833-839
Epstein, R. M. (2003). Mindful Practice in Action (II): Cultivating Habits of Mind. *Families, Systems & Health, 21*, 11-17
Epstein, R. M. & Hundert, E. M. (2002). Defining and Assessing Professional Competence. *The Journal of the American Medical Association, 287*(2), 226-235.
Eraut, M. (1994). Learning professional processes: public knowledge and personal experience. In M. Eraut (Ed.), *Developing professional knowledge and competence* (pp. 100-122). London: Falmer Press.
Erber, R., & Fiske, S. T. (1984). Outcome dependency and attention to inconsistent information. *Journal of Personality and Social Psychology, 47*(4), 709-726.
Ericsson, K. A. (2004). Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains. *Academic medicine, 79*(Supplement), S70-S81
Ericsson, K., & Simon, H. (1980). Verbal reports as data. *Psychological Review, 87*, 215-251.
Erwin, T. (1991). *Assessing Student Learning and Development*: Jossey-Bass.
Ertmer, P. A., & Newby, T. J. (1993). Behaviorism, Cognitivism, Constructivism: Comparing Critical Features From a Design Perspective. *Performance Improvement Quarterly, 6*(4), 50-72.
Ertmer, P. A., & Newby, T. J. (2013). Behaviorism, Cognitivism, Constructivism: Comparing Critical Features From an Instructional Design Perspective. *Performance Improvement Quarterly, 26*(2), 43-71.
Esses, V., Haddock, G., & Zanna, M. (1993). Values, stereotypes and emotions as determinants of intergroup attitudes. In D. Mackie & D. Hamilton (Eds.), *Affect, cognition, and stereotyping: Interactive processes in group perception* (pp. 137-166). San Diego: Academic Press.
Eva, K. W. (2001). Assessing Tutorial-Based Assessment. *Advances in Health Sciences Education, 6*(3), 243-257.
Eva, K. W. (2003). On the generality of specificity. *Medical education, 37*(7), 587-588.
Eva, K. W., O'Neill, P., Yeates, P., & Mann, K. (2012). Effect of Exposure to Good vs Poor Medical Trainee Performance on Attending Physician Ratings of Subsequent Performances. *JAMA, 308*(21), 2226-2232.
Ezzy, D. (2002). *Qualitative Analysis: Practice and innovation*. London: Routledge.
Faggiolani, C. (2011). Perceived Identity: applying Grounded Theory in Libraries. *JLIS.it, 2*(1).
Feldman, J. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*(2), 127-148.
Feldt, L. S., & Brennan, R. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York: Macmillan.

Ferris, G. R., Munyonm, T. P., Basik, K., & Buckley, M. R. (2008). The performance evaluation context: Social, emotional, cognitive, political, and relationship components. *Human Resource Management Review, 18*, 146-163

Fisher, C. D. (2008). What if we took within-person performance variability seriously? *Industrial and Organizational Psychology, 1*, 185-189

Fisher, C. D., & Noble, C. S. (2004). A Within-Person Examination of Correlates of Performance and Emotions While Working. *Human Performance, 17*(2), 145-168

Fiske, S. (1989). Examining the role of intent: Toward understanding its role in stereotyping and prejudice. In J. Uleman & J. Bargh (Eds.), *Unintended thought* (pp. 253-286). New York: Guilford Press.

Fiske, S., Lin, M., & Neuberg, S. (1990). The continuum model: Ten years later. In S. Chaiken & Y. Trope (Eds.), *Dual process theories in social psychology*. New York: Guuilford Press.

Fiske, S., & Neuberg, S. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1-74). San Diego, CA: Academic Press.

Fiske, S., & Pavelchak, M. (1986). Category-based versus piecemeal-based affective responses. In R. Sorrentino & E. Higgins (Eds.), *Handbook of motivation and cognition* (pp. 167-203). New York: Guilford Press.

Fiske, S. T. (1993). Social cognition and social perception. *Annual review of psychology, 44*(1), 155-194.

Fiske, S. T., & Cox, M. G. (1979). Person concepts: The effect of target familiarity and descriptive purpose on the process of describing others. *Journal of Personality, 47*(1), 136

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from the perceived status and competition. *Journal of Personality and Social Psychology, 82*, 878-902

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77-83

Fiske, S. T., Xu, J., Cuddy, A., & Glick, P. (1999). (Dis)respecting versus (dis)liking: Status and interdependence predict ambivalent stereotypes of competence and warmth. *Journal of Social Issues, 55*, 473–491

Forgas, J. P. (1994). The role of emotion in social judgments: An introductory review and an Affect Infusion Model (AIM). *European Journal of Social Psychology, 24*(1), 1-24.

Foster, S. L., & Cone, J. D. (1995). Validity Issues in Clinical Assessment. *Psychological Assessment, 7*(3), 248-260

Francis, J. J., Johnston, M., Robertson, C., Glidewell, L., Entwistle, V., Eccles, M. P., & Grimshaw, J. M. (2010). What is an adequate sample size? Operationalising data saturation for theory-based interview studies. *Psychology & Health, 25*(10), 1229.

Fraser, S. W., & Greenhalgh, T. (2001). Complexity Science: Coping With Complexity: Educating For Capability. *BMJ: British Medical Journal, 323*(7316), 799-803.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Frenk, J., Chen, L., Bhutta, Z. A., Cohen, J., Crisp, N., Evans, T., . . . Zurayk, H. (2010). Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *The Lancet, 376*(9756), 1923-1958.

Friedman Ben-David, M. (2000a). The role of assessment in expanding professional horizons. *Medical Teacher, 22*(5), 472-477.

Friedman Ben-David, M. (2000b). Standard setting in student assessment. *Medical Teacher, 22*(2), 120-130.

Fromme, H. B., Karani, R. & Downing, S. M. (2009). Direct observation in medical education: a review of the literature and evidence for validity. *Mount Sinai Journal of Medicine, 76*(4), 365-371.

Gallagher, T. J., Hartung, P. J., Gerzina, H., Gregory, S. W., Jr. & Merolla, D. (2005). Further analysis of a doctor-patient nonverbal communication instrument. *Patient Education & Counseling, 57*(3), 262-271.

Gilbert, D. T., & Hixon, J. G. (1991). The Trouble of Thinking: Activation and Application of Stereotypic Beliefs. *Journal of Personality and Social Psychology, 60*(4), 509-517.

Gilbert, D. T., & Jones, E. E. (1986). Perceiver-Induced Constraint: Interpretations of Self-Generated Reality. *Journal of Personality and Social Psychology, 50*(2), 269-280.

Gilbert, D. T., Jones, E. E., & Pelham, B. W. (1987). Influence and Inference: What the Active Perceiver Overlooks. *Journal of Personality and Social Psychology, 52*(5), 861-870.

Gilovich, T., & Inbar, Y. (2011). Angry (or disgusted), but adjusting? The effect of specific emotions on adjustment from self-generated anchors. *Social Psychological and Personality Science, 2*(6), 563-569.

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

Gingerich, A., Regehr, G., & Eva, K. W. (2011). Rater-based assessments as social judgments: rethinking the etiology of rater errors. *Academic medicine : journal of the Association of American Medical Colleges, 86*(10 Suppl), S1-S7.

Ginsburg, S., Bernabeo, E., Ross, K. M., & Holmboe, E. S. (2012). "It depends": results of a qualitative study investigating how practicing internists approach professional dilemmas. *Academic medicine : journal of the Association of American Medical Colleges, 87*(12), 1685-1693.

Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Academic medicine : journal of the Association of American Medical Colleges, 85*(5), 780-786.

Ginsburg, S., Regehr, G., Hatala, R., McNaughton, N., Frohna, A., Hodges, B., . . . Stern, D. (2000). Context, conflict, and resolution: a new conceptual framework for evaluating professionalism. *Academic medicine : journal of the Association of American Medical Colleges, 75*(10 Suppl), S6.

Gipps, C. (1994). Quality in teacher assessment. In W. Harlen (Ed.), *Enhancing quality in assessment*. London: Chapman.

Gipps, C. (1999). Socio-Cultural Aspects of Assessment. *Review of Research in Education, 24*, 355-392.

Given, L. M. (2008). *The Sage encyclopedia of qualitative research methods*. London;Los Angeles;: SAGE.

Glaser, B. G. (1978). *Theoretical sensitivity*. Mill Valley, Calif: Sociology Press.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: strategies for qualitative research*. New York: Aldine de Gruyter.

Gorawara-Bhat, R., Cook, M. A. & Sachs, G. A. (2007). Nonverbal communication in doctor-elderly patient transactions (NDEPT): development of a tool. *Patient Education & Counseling, 66*(2), 223-234.

Gordon, H. S., Street, R. L., Jr., Sharf, B. F. & Souchek, J. (2006). Racial differences in doctors' information-giving and patients' participation. *Cancer, 107*(6), 1313-1320.

Gormley, G. (2011). Summative OSCEs in undergraduate medical education. *The Ulster medical journal, 80*(3), 127-132

Govaerts, M. J., Schuwirth, L., Van der Vleuten, C., & Muijtjens, A. M. M. (2011). Workplace-based assessment: Effects of rater expertise. *advances in Health Sciences Education: Theory and Practice, 16*(2), 151-165.

Govaerts, M. J., van de Wiel, M. W., Schuwirth, L. W., Van der Vleuten, C. P., & Muijtjens, A. M. M. (2013). Workplace-based assessment: raters' performance theories and constructs. *Advances in Health Sciences Education, 18*(3), 375-396

Govaerts, M. J. B., van de Wiel, M. W. J., & van der Vleuten, C. P. M. (2013). Quality of feedback following performance assessments: does assessor expertise matter? *European Journal of Training and Development, 37*(1), 105-125.

Govaerts, M., & van der Vleuten, C. (2013). Validity in work-based assessment: expanding our horizons. *Medical education, 47*, 1164-1174

Govaerts, M. J. B., van der Vleuten, C. P. M., Schuwirth, L. W. T., & Muijtjens, A. M. M. (2007). Broadening Perspectives on Clinical Performance Assessment: Rethinking the Nature of In-training Assessment. *Advances in Health Sciences Education, 12*(2), 239-260.

Grand'Maison, P., Brailovsky, C. A., & Lescop, J. (1996). Content validity of the Quebec licensing examination (OSCE). Assessed by practising physicians. *Canadian family physician Médecin de famille canadien, 42*, 254-259

Gray, J. D. (1996). Global rating scales in residency education. *Academic medicine : journal of the Association of American Medical Colleges, 71*(1 Suppl), S55-63.

Green, M. L., & Holmboe, E. (2010). Perspective: the ACGME toolbox: half empty or half full? *Academic medicine : journal of the Association of American Medical Colleges, 85*(5), 787-790.

Green, S. (2002). *CRITERION REFERENCED ASSESSMENT AS A GUIDE TO LEARNING THE IMPORTANCE OF PROGRESSION AND RELIABILITY*. Paper presented at the Association for the Study of Evaluation in Education in Southern Africa International Conference.

Greifeneder, R., & Bless, H. (2010). The fate of activated information in impression formation: fluency of concept activation moderates the emergence of assimilation versus contrast. *The British journal of social psychology / the British Psychological Society, 49*(Pt 2), 405.

Griffith, C. H., 3rd, Wilson, J. F., Langer, S. & Haist, S. A. (2003). House staff nonverbal communication skills and standardized patient satisfaction. *Journal of General Internal Medicine, 18*(3), 170-174.

Gronlund, N. (1976). *Measuring and evaluating in teaching* (3rd ed.). London: Collier Macmillan.

Gruppen, L., & Frohna, A. (2002). Clinical reasoning. In G. Norman, C. van der Vleuten & D. Newble (Eds.), *International Handbook of Research in Medical Education* (pp. 205-230). Dordrecht: Kluwer Academic Publishers.

Guba, E. (1990). The alternative paradigm dialog. In E. Guba (Ed.), *The Paradigm Dialog* (pp. 17-30). Newbury Park, CA: Sage Publications.

Guba, E., & Lincoln, Y. (1994). Competing paradigms in qualitative research. In N. Denzin & Y. Lincoln (Eds.), *Handbook of Qualitative Research* (pp. 105-117). Thousand Oaks, CA: Sage Publications.

Guba, E. G., & Lincoln, Y. S. (1982). Epistemological and Methodological Bases of Naturalistic Inquiry. *Educational Communication and Technology, 30*(4), 233-252.

Guilford, J. P. (1967). *The Nature of Human Intelligence*. New York: McGraw-Hill.

Hager, P. (2011). *Theories of Workplace Learning*. Los Angeles, CA: Sage Publications.

Hager, P., & Hodkinson, P. (2009). Moving beyond the metaphor of transfer of learning. *British Educational Research Journal, 35*(4), 619-638

Hall, J., Harrigan, J. & Rosenthal, R. (1995). Non-verbal behaviour in clinical-patient interaction. *Applied & Preventive Psychology, 4*, 21-37.

Hall, J. A., Roter, D. L. & Rand, C. S. (1981). Communication of affect between patient and physician. *Journal of Health & Social Behavior, 22*(1), 18-30.

Hall, P., Keely, E., Dojeiji, S., Byszewski, A. & Marks, M. (2004). Communication skills, cultural challenges and individual support: challenges of international medical graduates in a Canadian healthcare environment. *Medical Teacher, 26*(2), 120-125.

Hamdy, H., Telmesani, A., Wardy, N., Abdel-Khalek, N., Carruthers, G., Hassan, F., . . . O'malley, K. (2010). Undergraduate medical education in the Gulf Cooperation Council: A multi-countries study (Part 2). *Medical Teacher, 32*, 290–295.

Hamilton, D. (1979). A cognitive-attributional analysis of stereotyping. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (pp. 53-84). San Diego, CA: Academi Press.

Hamilton, D., & Trolier, T. (1986). Stereotypes and stereotyping: An overview of the cognitive approach. In J. Dovidio & S. Gaertner (Eds.), *Prejudice, discrimination, and racism* (pp. 127-163). San Diego, CA: Academic Press.

Hamilton, D. L., Driscoll, D. M., & Worth, L. T. (1989). Cognitive Organization of Impressions: Effects of Incongruency in Complex Representations. *Journal of Personality and Social Psychology, 57*(6), 925-939.

Hamilton, D. L., Sherman, S. J., & Ruvolo, C. M. (1990). Stereotype-Based Expectancies: Effects on Information Processing and Social Behavior. *Journal of Social Issues, 46*(2), 35-60.

Hanson, F. A. (1993). *Testing Testing: social consequences of the examined life*. Berkeley, California: University of California Press.

Harasym, P. H., Woloschuk, W., & Cunning, L. (2008). Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Advances in Health Sciences Education, 13*(5), 617-632.

Harden, R. (1988). What is an OSCE? *Medical Teacher, 10*(1), 19-22

Harden, R. & Gleeson, F. (1979). Assessment of clinical competence using an objective structured clinical examination. *medical Education, 13*, 41-47.

Harden, R., lilley , P., & Patricio, M. (2015). *THE DEFINITIVE GUIDE TO THE OSCE: The Objective Structured Clinical Examination as a performance assessment* (1st ed.). Edinburgh: Elsevier.

Harden, R. M. (2014). Progression in competency-based education. *Medical education, 48*(8), 838-838.

Harden, R. M., Hart, I. R., Mulholland, H., Association for the Study of Medical, E., & University of Dundee. Centre for Medical, E. (1992). *Approaches to the assessment of clinical competence: Part 1*. Dundee: Centre for Medical Education.

Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). Assessment Of Clinical Competence Using Objective Structured Examination. *The British Medical Journal, 1*(5955), 447-451.

Harris, J. (2002). The Correspondence Method as a Data-gathering Technique in Qualitative Enquiry. *International Journal of Qualitative Methods, 1*(4), 1

Hatala, R., Ainslie, M., Kassen, B. O., Mackie, I., & Roberts, J. M. (2006). Assessing the mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Medical education, 40*(10), 950-956.

Hatala, R., & Norman, G. R. (1999). In-training evaluation during an internal medicine clerkship. *Academic medicine : journal of the Association of American Medical Colleges, 74*(10 Suppl), S118-120.

Hattie, J. & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Hauer, K. E., Holmboe, E. S., & Kogan, J. R. (2009). Tools for Direct Observation and Assessment of Clinical Skills of Medical Trainees: A Systematic Review. *JAMA, 302*(12), 1316-1326.

Hawkins, R. E., Margolis, M. J., Durning, S. J., & Norcini, J. J. (2010). Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. *Academic medicine : journal of the Association of American Medical Colleges, 85*(9), 1453-1461.

Hays, R. (2008). Assessment in medical education: roles for clinical teachers. *The Clinical Teacher, 5*(1), 23-27

Herbers, J. J. E., Noel, G. L., Cooper, G. S., Harvey, J., Pangaro, L. N., & Weaver, M. J. (1989). How accurate are faculty evaluations of clinical competence? *Journal of General Internal Medicine, 4*(3), 202-208.

Hill, F., Kendall, K., Galbraith, K., & Crossley, J. (2009). Implementing the undergraduate mini-CEX: a tailored approach at Southampton University. *Medical education, 43*(4), 326-334.

Hodder, R. V., Rivington, R. N., Calcutt, L. E., & Hart, I. R. (1989). The effectiveness of immediate feedback during the objective structured clinical examination. *Medical education, 23*(2), 184

Hodges, B. (2006). Medical education and the maintenance of incompetence. *Medical Teacher, 28*(8), 690-696.
Hodges, B. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Medical Teacher, 35*(7), 564-568.
Hodges, B. D. (2010). A tea-steeping or i-Doc model for medical education? *Academic medicine : journal of the Association of American Medical Colleges, 85*(9 Suppl), S34-S44.
Hodges, B., & McIlroy, J. H. (2003). Analytical global OSCE ratings are sensitive to level of training. *Medical education, 37*(11), 1012-1016
Hodges, B., Regehr, G., Hanson, M. & McNaughton, N. (1997). An objective structured clinical examination for evaluating psychiatric clinical clerks. *Academic Medicine, 72*(8), 715-721.
Hodges, B., Turnbull, J., Cohen, R., Bienenstock, A., & Norman, G. (1996). Evaluating communication skills in the objective structured clinical examination format: reliability and generalizability. *Medical education, 30*(1), 38-43
Hoffman, B., Lance, C. E., Bynum, B., & Gentry, W. A. (2010). RATER SOURCE EFFECTS ARE ALIVE AND WELL AFTER ALL. *Personnel Psychology, 63*(1), 119-151.
Hofstadter, D. (2007). *I am a strange loop*. New York, NY: Basic Book.
Holmboe, E. S., Hawkins, R. E., & Huot, S. J. (2004). Effects of training in direct observation of medical residents' clinical competence: a randomized trial. *Annals of internal medicine, 140*(11), 874.
Holmboe, E. S., Huot, S., Chung, J., Norcini, J., & Hawkins, R. E. (2003). Construct validity of the miniclinical evaluation exercise (miniCEX). *Academic medicine : journal of the Association of American Medical Colleges, 78*(8), 826-830.
Holmboe, E. S., Sherbino, J., Long, D., Swing, S., & Frank, J. (2010). The role of assessment in competency-based medical education. *Medical Teacher, 32*, 676-682.
Holmboe, E. S., Ward, D. S., Reznick, R. K., Katsufrakis, P. J., Leslie, K. M., Patel, V. L., . . . Nelson, E. A. (2011). Faculty development in assessment: the missing link in competency-based medical education. *Academic medicine, 86*(4), 460.
Holmboe, E. S., Yepes, M., Williams, F., & Huot, S. J. (2004). Feedback and the mini clinical evaluation exercise. *Journal of General Internal Medicine, 19*(S5), 558-561.
Holte, A. (1990). Professional communication skills. *Scandinavian Journal of Primary Health Care, 8*(3), 131-132.
Humphrey-Murto, S., Touchie, C., & Smee, S. (2013). Objective structured Clinical Examinations. In K. Walsh (Ed.), *Oxford Textbook of Medical Education* (pp. 524-536). Oxford: Oxford University Press.
Hurlbert, R. T. (2006). *Comprehending Behavioral Statistics*. Toronto, Ontario, Canada: Thomson Wadsworth.
Ilgen, D. R., Barnes-Farrell, J. L., & McKellin, D. B. (1993). Performance Appraisal Process Research in the 1980s: What Has It Contributed to Appraisals in Use? *Organizational Behavior and Human Decision Processes, 54*(3), 321-368.
Illing, J. (2010). Thinking about research: theoretical perspectives, ethics and scholarship. In T. Swanwick (Ed.), *Understanding Medical Education: Evidence, Theory and Practice* (pp. 283-300). Oxford: Wiley-Blackwell.
Inbar, Y., & Gilovich, T. (2011). Angry (or disgusted), but adjusting? The effect of specific emotions on adjustment from self-generated anchors. *Social Psychological & Personality Science, 2*, 563–569
Institue of Medicine. (2001) *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: Institute of Medicine.
Irby, D. M., Cooke, M., & O'Brien, B. C. (2010). Calls for reform of medical education by the Carnegie Foundation for the Advancement of Teaching: 1910 and 2010. *Academic medicine : journal of the Association of American Medical Colleges, 85*(2), 220-227.

Ishikawa, H., Hashimoto, H., Kinoshita, M., Fujimori, S., Shimizu, T. & Yano, E. (2006). Evaluating medical students' non-verbal communication during the objective structured clinical examination. *Medical Education, 40*(12), 1180-1187.

Jackson, L. A., Sullivan, L. A., & Hodge, C. N. (1993). Stereotype Effects on Attributions, Predictions, and Evaluations: No Two Social Judgments Are Quite Alike. *Journal of Personality and Social Psychology, 65*(1), 69-84.

Jackson, N., Jamieson, A., & Khan, A. (2007). *Assessment in medical education and training: a practical guide*. Abingdon: Radcliffe.

Jacowitz, K. E., & Kahneman, D. (1995). Measures of Anchoring in Estimation Tasks. *Personality and Social Psychology Bulletin, 21*(11), 1161-1166.

Johnston, B. (2004). Summative assessment of portfolios: an examination of different approaches to agreement over outcomes. *Studies in Higher Education, 29*(3), 395–412

Jones, R. (1995). Why do qualitative research? *British Medical Journal, 311*, 2.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental Dimensions of Social Judgment: Understanding the Relations Between Judgments of Competence and Warmth. *Journal of Personality and Social Psychology, 89*(6), 899-913

Jussim, L., Coleman, L. M., & Lerch, L. (1987). The Nature of Stereotypes: A Comparison and Integration of Three Theories. *Journal of Personality and Social Psychology, 52*(3), 536-546.

Jussim, L., Nelson, T. E., Manis, M., & Soffin, S. (1995). Prejudice, Stereotypes, and Labeling Effects: Sources of Bias in Person Perception. *Journal of Personality and Social Psychology, 68*(2), 228-246.

Kane, M. T. (2008). Terminology, Emphasis, and Utility in Validation. *Educational Researcher, 37*(2), 76-82

Karlssona, G., & Tham, K. (2006). Correlating facts or interpreting meaning: Two different epistemological projects within medical research. *Scandinavian Journal of Occupational Therapy, 13*(2), 68-75

Kassebaum, D. G., & Eaglen, R. H. (1999). Shortcomings in the evaluation of students' clinical skills and behaviors in medical school. *Academic medicine : journal of the Association of American Medical Colleges, 74*(7), 842-849.

Keller, L. A., Mazor, K. M., Swaminathan, H., & Pugnaire, M. P. (2000). An Investigation of the Impacts of Different Generalizability Study Designs on Estimates of Variance Components and Generalizability Coefficients. *Academic medicine, 75*(Supplement), S21-S24.

Kennedy, T. J. T., & Lingard, L. A. (2006). Making sense of grounded theory in medical education. *Medical education, 40*(2), 101-108.

Kenny, D. (1994). *Interpersonal Perception: A social Relations Analysis*. New York, NY: Guilford Press.

Kenny, D. A. (1988). Interpersonal Perception: A Social Relations Analysis. *Journal of Social and Personal Relationships, 5*(2), 247-261.

Kenny, D. A. (2004). PERSON: A General Model of Interpersonal Perception. *Personality and Social Psychology Review, 8*(3), 265-280.

Khan, K., Ramachandran, S., Gaunt, K., & Pushkar, P. (2013). The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part I: an historical and theoritical perspective. *Medical Teacher, 35*(9), e1437-1446.

Kim, K. S. (2010). Introduction and Administration of the Clinical Skill Test of the Medical Licensing Examination, Republic of Korea (2009). *Journal of Educational Evaluation for Health Professions, 7*, 4

Kirkpatrick, D. & Kirkpatrick, J. (2005). *Evaluating training programs: The four levels*. San Francisco: Berrett-Koehler.

Klass, D. (2000). Reevaluation of clinical competency. *American Journal of Physical Medicine & Rehabilitation, 79*, 481-486.

Klein, G. (2009). *Streetlights and Shadows: Searching for the Keys to Adaptive Decision Making*. Cambridge, MA: MIT Press.

Klein, K., & Creech, B. (1982). Race, Rape, and Bias: Distortion of Prior Odds and Meaning Changes. *Basic and Applied Social Psychology, 3*(1), 21-33.
Klein, W. M., & Kunda, Z. (1992). Motivated person perception: Constructing justifications for desired beliefs. *Journal of Experimental Social Psychology, 28*(2), 145-168
Klimoski, R., & Donahue, L. (2001). Person perception in organizations: An overview of the field. In M. London (Ed.), *How People Evaluate Others in Organizations* (pp. 5-43). Mahwah, NJ: Lawrence Erlbaum Associate, Inc.
Klimoski, R., & Inks, L. (1990). Accountability forces in performance appraisal. *Organizational Behavior and Human Decision Processes, 45*, 194-208.
Kluger, A., & van Dijk, D. (2010). Feedback, the various tasks of the doctor, and the feedforward alternative. *Medical education, 44*, 1166-1174.
Kneebone, R. (2003). Simulation in surgical training: educational issues and practical implications. *Medical education, 37*(3), 267-277
Koch, M. J., & DeLuca, C. (2012). Rethinking validation in complex high-stakes assessment contexts. *Assessment in Education: Principles, Policy & Practice, 19*(1), 99-116
Kogan, J. R., Conforti, L., Bernabeo, E., Iobst, W. & Holmboe, E. (2011). Opening the black box of clinical skills assessment via observation: a conceptual model. *Medical Education, 45*(10), 1048-1060.
Kogan, J., holmboe, E., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A system review. *JAMA: The Journal of the American Medical Association, 302*(12), 1316-1326.
Kogan, J. R., Hess, B. J., Conforti, L. N., & Holmboe, E. S. (2010). What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Academic medicine : journal of the Association of American Medical Colleges, 85*(10 Suppl), S25-S28.
Koriat, A. (1993). How Do We Know That We Know? The Accessibility Model of the Feeling of Knowing. *Psychological Review, 100*(4), 609-639.
Kreiter, C. D., & Ferguson, K. J. (2001). Examining the Generalizability of Ratings across Clerkships Using a Clinical Evaluation Form. *Evaluation & the Health Professions, 24*(1), 36-46
Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology, 19*(5), 448-468.
Kunda, Z., Miller, D. T., & Claire, T. (1990). Combining social concepts: The role of causal reasoning. *Cognitive Science, 14*(4), 551-577.
Kunda, Z., & Sherman-Williams, B. (1993). Stereotypes and the Construal of Individuating Information. *Personality and Social Psychology Bulletin, 19*(1), 90-99.
Kunda, Z., Sinclair, L., & Griffin, D. (1995). *Equal ratings but seperate meanings: Stereotypes and the construal of traits*. Unpublished manuscript.
Kunda, Z., & Spencer, S. J. (2003). When do stereotypes come to mind and when do they color judgment? A goal-based theoretical framework for stereotype activation and application. *Psychological Bulletin, 129*(4), 522.
Kunda, Z., & Thagard, P. (1996). Forming Impressions From Stereotypes, Traits, and Behaviors: A Parallel-Constraint-Satisfaction Theory. *Psychological Review, 103*(2), 284-308.
Kuper, A., Reeves, S., Albert, M., & Hodges, B. D. (2007). Assessment: do we need to broaden our methodological horizons? *Medical education, 41*(12), 1121-1123.
Laidlaw, A. & Hart, J. (2011). Communication skills: an essential component of medical curricula. Part I: Assessment of clinical communication: AMEE Guide No. 51. *Medical Teacher, 33*(1), 6-8.
Laing, R. D., Phillipson, H., & Lea, A. R. (1966). *Interpersonal perception: a theory and a method of research*. London: Tavistock Publications.
Lance, C. E., Hoffman, B. J., Gentry, W. A., & Baranik, L. E. (2008). Rater source factors represent important subcomponents of the criterion construct space, not rater bias. *Human Resource Management Review, 18*(4), 223-232

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.
Larsen, K. & Smith, C. (1981). Assessment of non-verbal communication in the patient-physician interview. *Journal of Family Practice, 12*, 481-488.
Leach, D. C. (2002). Competence Is a Habit. *JAMA, 287*(2), 243-244.
LeDoux, J. E. (1999). *The emotional brain: the mysterious underpinnings of emotional life*. London: Phoenix.
Lejk, M. & Wyvill, M. (2001). Peer assessment of contributions to a group project: a comparison of holistic and category based approaches. *Assessment and Evaluation in Higher Education, 26*, 61-72.
Leung, W.-C. (2002). Competency Based Medical Training: Review. *BMJ: British Medical Journal, 325*(7366), 693-695
Levy, P. E., & Williams, J. R. (2004). The Social Context of Performance Appraisal: A Review and Framework for the Future. *Journal of Management, 30*(6), 881-905.
Leyens, J.-P., & Fiske, S. (1994). Impression formation: From recitals to symphonie fantastique. In P. Devine, D. Hamilton & T. Ostrom (Eds.), *Social Cognition: Impact on Social Psychology* (pp. 39-75). San Diego, Calif: Academic Press.
Lingard, L. (2007). Qualitative research in the RIME community: critical reflections and future directions. *Academic medicine : journal of the Association of American Medical Colleges, 82*(10 Suppl), S129-S130.
Lingard, L. (2009). What we see and don't see when we look at 'competence': notes on a god term. *Advances in health sciences education : theory and practice, 14*(5), 625.
Lingard, L., & Kennedy, T. J. (2010). Qualitative research methods in medical education. In T. Swanwick (Ed.), *Understanding Medical Education: Evidence, Theory and Practice* (pp. 323-335). Oxford: Wiley-Blackwell.
Lingard, L., Vleuten, C. P. M. v. d., Watling, C., & essen, E. (2012). Learning from clinical work: the roles of learning cues and credibility judgements. *Medical education, 46*(2), 192-200.
Lingle, J. H. (1979). Thematic effects of person judgments on impression organization. *Journal of Personality and Social Psychology, 37*(5), 674-687.
Linville, P. W., & Jones, E. E. (1980). Polarized appraisals of out-group members. *Journal of Personality and Social Psychology, 38*(5), 689-703.
Lippman, W. (1922). *Public opinion*. New York: Harcourt & Brace.
Littlefield, J. H., Darosa, D. A., Paukert, J., Williams, R. G., Klamen, D. L., & Schoolfield, J. D. (2005). Improving resident performance assessment data: numeric precision and narrative specificity. *Academic medicine : journal of the Association of American Medical Colleges, 80*(5), 489-495.
Locksley, A., Borgida, E., Brekke, N., & Hepburn, C. (1980). Sex stereotypes and social judgment. *Journal of Personality and Social Psychology, 39*(5), 821-831.
Locksley, A., Hepburn, C., & Ortiz, V. (1982). Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of Experimental Social Psychology, 18*(1), 23-42.
Lowry, S. (1993). *Medical Education*. London: BMJ books.
Lurie, S. J., Mooney, C. J., & lyness, J. M. (2009a). In reply: Letters to editor: How should the ACGME core competencies be measured? *Academic medicine, 84*, 1173.
Lurie, S. J., Mooney, C. J., & Lyness, J. M. (2009b). Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. *Academic medicine : journal of the Association of American Medical Colleges, 84*(3), 301-309.
Lurie, S. J., Mooney, C. J., & Lyness, J. M. (2011). Commentary: pitfalls in assessment of competency-based educational objectives. *Academic medicine, 86*(4), 412.
MacLellan, A.-M., Brailovsky, C., Rainsberry, P., Bowmer, I., & Desrochers, M. (2010). Examination outcomes for international medical graduates pursuing or completing family medicine residency training in Quebec. *Canadian Family Physician, 56*(9), 912-918.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annual review of psychology, 51*(1), 93-120.
Macrae, C. N., & Bodenhausen, G. V. (2001). Social cognition: categorical person perception. *British journal of psychology, 92*(Pt 1), 239-255.
Macrae, C. N., Bodenhausen, G. V., & Milne, A. B. (1995). The Dissection of Selection in Person Perception: Inhibitory Processes in Social Stereotyping. *Journal of Personality and Social Psychology, 69*(3), 397-407.
Macrae, C. N., Bodenhausen, G. V., Milne, A. B., & Jetten, J. (1994). Out of Mind but Back in Sight: Stereotypes on the Rebound. *Journal of Personality and Social Psychology, 67*(5), 808-817.
Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as Energy-Saving Devices: A Peek Inside the Cognitive Toolbox. *Journal of Personality and Social Psychology, 66*(1), 37-47.
MacRae, C. N., Schloerscheidt, A. M., Bodenhausen, G. V., & Milne, A. B. (2002). Creating memory illusions: Expectancy-based processing and the generation of false memories. *Memory, 10*(1), 63-80.
Malle, B. F., & Pearce, G. E. (2001). Attention to Behavioral Events During Interaction: Two Actor-Observer Gaps and Three Attempts to Close Them. *Journal of Personality and Social Psychology, 81*(2), 278-294.
Malloy, T. E., & Albright, L. (1990). Interpersonal Perception in a Social Context. *Journal of Personality and Social Psychology, 58*(3), 419-428
Mann, K. V. (2011). Theoretical perspectives in medical education: past experience and future possibilities. *Medical education, 45*(1), 60-68
Margolis, M. J., Clauser, B. E., Cuddy, M. M., Ciccone, A., Mee, J., Harik, P., & Hawkins, R. E. (2006). Use of the Mini-Clinical Evaluation Exercise to Rate Examinee Performance on a Multiple-Station Clinical Skills Examination: A Validity Study. *Academic medicine, 81*(Suppl), S56-S60.
Marshall, M. (1996). Sampling for qualitative research. *Family Practice, 13*, 522-525
Marshall, C., & Rossman, G. B. (1999). *Designing qualitative research* (Vol. 3rd). Thousand Oaks, CA;London;: Sage Publications.
Marshall, V., & Ludbrook, J. (1972). The relative importance of patient and examiner variability in a test of clinical skills. *British journal of Medical Education, 6*, 212-217.
Martin, P. Y., & Turner, B. A. (1986). Grounded Theory and Organizational Research. *The Journal of Applied Behavioral Science, 22*(2), 141-157.
Mays, N. & Pope, C. (1995). Rigour and qualitative research. *British Medical Journal, 311*, 109-112.
Mazor, K. M., Zanetti, M. L., Alper, E. J., Hatem, D., Barrett, S. V., Meterko, V., . . . Pugnaire, M. P. (2007). Assessing professionalism in the context of an objective structured clinical examination: an in-depth study of the rating process. *Medical education, 41*(4), 331-340.
McAleer, S. (2001). Formative and summative assessment. In J. A. Dent & R. M. Harden (Eds.), *A practical Guide for Medical Teachers*. London: Churchill Livingstone.
McArthur, L. Z., & Baron, R. M. (1983). Toward an ecological theory of social perception. *Psychological Review, 90*(3), 215-238.
McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399-414
McFarlane, D. (2004). Not everythig that counts can be counted. Retrieved 15 June, 2014, from http://www.teachers.ab.ca/Publications/The%20Learning%20Team/Volume%208/Number%203/Pages/Not%20everything%20that%20counts%20can%20be%20counted.aspx
McFaul, P. B., Taylor, D. J., & Howie, P. W. (1993). The assessment of clinical competence in obstetrics and gynaecology in two medical schools by an objective structured clinical examination. *British journal of obstetrics and gynaecology, 100*(9), 842-846
McGaghie, W. C., Butter, J., & Kaye, M. (2009). Observational assessment. In S. Downing & R. Yudkowsky (Eds.), *Assessment in health professions education* (pp. 185-216). New York, NY: Routledge.

McGaghie, W. C., & Lipson, L. (1978). *Competency-based Curriculum Development in Medical Education: An Introduction*. Geneva: World Health Organization.
McLeod, J. M., & Chaffee, S. H. (1973). Interpersonal Approaches to Communication Research. *American Behavioral Scientist, 16*(4), 469-499
McManus, I. C., Elder, A. T., & Dacre, J. (2013). Investigating possible ethnicity and sex bias in clinical examiners: an analysis of data from the MRCP(UK) PACES and nPACES examinations. *BMC medical education, 13*(1), 103-103.
McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC medical education, 6*(1), 42-42.
McManus, I. C., Woolf, K., & Dacre, J. (2008). The educational background and qualifications of UK medical students from ethnic minorities. *BMC medical education, 8*(1), 21-21.
Medical Council of Canada. (2011). Evidence for the Validity of a Clinical Skill Assessment. Retrieved 20 September, 2015, from http://mcc.ca/wp-content/uploads/technical-reports-wood-2011.pdf
Mennin, S. (2010). Teaching, Learning, Complexity and Health Professions Education (Vol. 20, pp. 162-165): International Association of Medical Science Educators (IAMSEI).
Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.). Washington DC: Oryx Press.
Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic medicine : journal of the Association of American Medical Colleges, 65*(9 Suppl), S63-67.
Mischel, W., & Shoda, Y. (1995). A Cognitive-Affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure. *Psychological Review, 102*(2), 246-268.
Mitchum, A. L., & Kelley, C. M. (2010). Solve the Problem First: Constructive Solution Strategies Can Influence the Accuracy of Retrospective Confidence Judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 699-710.
Mohr, C. D., & 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.
Monteith, M. J., Sherman, J. W., & Devine, P. G. (1998). Suppression as a stereotype control strategy. *Personality and Social Psychology Review, 2*, 63–82
Moskowitz, G. B., & Li, P. (2011). Egalitarian goals trigger stereotype inhibition: A proactive form of stereotype control. *Journal of Experimental Social Psychology, 47*(1), 103-116.
Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.
Moss, P. A. (1996). Enlarging the dialogue in educational measurement: voices from interpretive research traditions. *Educational Research, 25*(1), 20-28
Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Chapter 4: Validity in Educational Assessment. *Review of Research in Education, 30*(1), 109-162
Murphy, J. F. A. (2007). Assessment in medical education. *Irish medical journal, 100*(2), 356.
Murphy, K., & Cleveland, J. (1995). *Understanding performance appraisal: Social organizational and goal-based perspectives*. Thousand Oaks, CA: Sage Publications.
Murphy, K. R., Balzer, W. K., Lockhart, M. C., & Eisenman, E. J. (1985). Effects of previous performance on evaluations of present performance. *Journal of Applied Psychology, 70*(1), 72-84.
Murphy, K. R., Cleveland, J. N., Skattebo, A. L., & Kinney, T. B. (2004). Raters Who Pursue Different Goals Give Different Ratings. *Journal of Applied Psychology, 89*(1), 158-164.
Mussweiler, T. (2001a). Focus of Comparison as a Determinant of Assimilation Versus Contrast in Social Comparison. *Personality and Social Psychology Bulletin, 27*(1), 38-47.
Mussweiler, T. (2001b). 'Seek and ye shall find': antecedents of assimilation and contrast in social comparison. *European Journal of Social Psychology, 31*(5), 499-509.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review, 110*(3), 472-489.

Nasca, T. J., Gonnella, J. S., Hojat, M., Veloskia, J., Erdmanna, J. B., Robesona, M., . . . Callahana, C. (2002). Conceptualization and measurement of clinical competence of residents: A brief rating form and its psychometric properties. *Medical Teacher, 24*(3), 299 –303

Neuberg, S. L., & Fiske, S. T. (1987). Motivational Influences on Impression Formation: Outcome Dependency, Accuracy-Driven Attention, and Individuating Processes. *Journal of Personality and Social Psychology, 53*(3), 431-444.

Newble, D. (1998). Assessment. In B. Jolly & L. Rees (Eds.), *Medical Education in the Millennium* (pp. 131-142). Oxford, UK: Oxford University Press.

Newble, D. (2004). The metric of medical education techniques for measuring clinical competence: Objective Structured clinical examinations. *Medical education, 44*, 199-203.

Newble, D. I., Hoare, J., & Sheldrake, P. F. (1980). The selection and training of examiners for clinical examinations. *Medical education, 14*(5), 345-349.

Newble, D. & Jaeger, K. (1983). The effect of assessments and examinations on the learning of medical students. *Medical Education, 17*, 165-171.

Nisbett, R., & Ross, L. (1980). *Human inference: strategies and shortcomings of social judgement*. Englewood Cliffs;London;: Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Noel, G. L., Herbers, J. E., Jr., Caplow, M. P., Cooper, G. S., Pangaro, L. N. & Harvey, J. (1992). How well do internal medicine faculty members evaluate the clinical skills of residents? *Annals of Internal Medicine, 117*(9), 757-765.

Norcini, J. (2003). ABC of teaching and learning in medicine work based assessment *British Medical journal, 326*, 753-755.

Norcini, J. (2003). Setting standard on educational test. *Medical Education, 37*, 464-469.

Norcini, J., Blamk, L., Arnold, G., & Kimball, H. (1997). Examiner differences in the mini-CEX. *advances in Health Sciences Education: Theory and Practice, 2*(1), 27-33.

Norcini, J., Anderson, B. B., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R., . . . Roberts, T. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 conference. *Medical Teacher, 33*(3), 206–214

Norcini, J., & Boulet, J. (2003). Methodological issues in the use of standardized patients for assessment. *Teaching and learning in medicine, 15*(4), 293

Norcini, J. & Burch, V. (2007). Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Medical Teacher, 29*(9), 855-871.

Norcini, J. J. (2005). Current perspectives in assessment: the assessment of performance at work. *Medical education, 39*(9), 880-889.

Norcini, J. J. (1999). Measurement issues in the use of simulation for testing professionals: Test development, test scoring, standard setting. In A. Tekian, C. McGuire & W. McGaghie (Eds.), *Innovative simulations for assessing professional competence*. Chicago, IL: University of Illinois.

Norcini, J. J., & McKinley, D. W. (2007). Assessment methods in medical education. *Teaching and Teacher Education, 23*(3), 239-250.

Norcini, J. J., Swanson, D. B., Grosso, L. J. & Webster, G. D. (1985). Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. *Medical Education, 19*(3), 238-247.

Norman, G. (2003). Postgraduate assessment - reliability and validity. *Journal of the Colleges of Medicine of South Africa, 47*, 71-75.

Norman, G. (2002). The long case versus objective structured clinical examinations. *British medical journal (Clinical research ed.), 324*(7340), 748-749

Norman, G. R., Tugwell, P., & Feightner, J. W. (1982). A comparison of resident performance on real and simulated patients. *Academic medicine, 57*(9), 708-715

Norman, G. R., Tugwell, P., Feightner, J. W., Muzzin, L. J., & Jacoby, L. L. (1985). Knowledge and clinical problem-solving. *Medical education, 19*(5), 344-356.

Northcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39*(1), 84-97.

Olson, G. M., Duffy, S. A., & Mack, R. L. (1984). Thinking out loud as a method of studying real time comprehension processes. In D. E. Kieras & M. A. Just (Eds.), *New methods in reading comprehension research* (pp. 253-286). Hillsdale: Erlbaum.

Ostroff, C., & Ilgen, D. R. (1992). Cognitive Categories of Raters and Rating Accuracy. *Journal of Business and Psychology, 7*(1), 3-26.

Paauw, D. S., Wenrich, M. D., Curtis, J. R., Carline, J. D., & Ramsey, P. G. (1995). Ability of primary care physicians to recognise physical findings associated with HIV infection. *JAMA, 274*(17), 1380–1382

Pangaro, L. N., & ten Cate, O. (2013). Frameworks for learner assessment in medicine: AMEE Guide no. 78. *Medical Teacher, 35*(6), e1197-1210.

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*(6), 407-418.

Parducci, A., & Perrett, L. F. (1971). Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology, 89*(2), 427-452.

Park, B. (1986). A Method for Studying the Development of Impressions of Real People. *Journal of Personality and Social Psychology, 51*(5), 907-917.

Park, B., DeKay, M. L., & Kraus, S. (1994). Aggregating Social Behavior Into Person Models: Perceiver-Induced Consistency. *Journal of Personality and Social Psychology, 66*(3), 437-459.

Park, B., & Judd, C. M. (1989). Agreement on Initial Impressions: Differences Due to Perceivers, Trait Dimensions, and Target Behaviors. *Journal of Personality and Social Psychology, 56*(4), 493-505.

Park, B., Kraus, S., & Ryan, C. S. (1997). Longitudinal Changes in Consensus as a Function of Acquaintance and Agreement in Liking. *Journal of Personality and Social Psychology, 72*(3), 604-616.

Patrício, M. F., Julião, M., Fareleira, F., & Carneiro, A. V. (2013). Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Medical Teacher, 35*(6), 503-514

Pelgrim, E. A. M., Kramer, A. W. M., Mokkink, H. G. A., van den Elsen, L., Grol, R. P. T. M., & Van der Vleuten, C. (2011). In-trining assessment using direct observation of single-patient encounters: A literature review. *advances in Health Sciences Education: Theory and Practice, 16*(1), 131-142.

Pell, G., Fuller, R., Homer, M., & Roberts, T. (2012). Is short term remediation after OSCE failure sustained? A retrospective analysis of the longitudinal attainment of underperforming students in OSCE assessments. *Medical Teacher, 34*(2), 146-150

Pendry, L. F., & Macrae, C. N. (1994). Stereotypes and Mental Life: The Case of the Motivated but Thwarted Tactician. *Journal of Experimental Social Psychology, 30*(4), 303-325.

Pennington, D. C. (2000). *Social cognition*. London: Routledge.

Pierre, R. B., Wierenga, A., Barton, M., Branday, J. M., & Christie, C. D. C. (2004). Student evaluation of an OSCE in paediatrics at the University of the West Indies, Jamaica. *BMC medical education, 4*(1), 22-22

Pope, C. & Mays, N. (1995). Reaching the parts other methods cannot reach: an introduction to qualitative methods in health services research. *British Medical Journal, 311*, 42-45.

Popham, W. J. (1980). Domain specification strategies. In R. A. Berk (Ed.), *Criterionreferenced measurement: the state of the art* (pp. 15-31). Baltimore and London: John: Hopkins University Press.

Pratto, F., & Bargh, J. A. (1991). Stereotyping based on apparently individuating information: Trait and global components of sex stereotypes under attention overload. *Journal of Experimental Social Psychology, 27*(1), 26-47.

Prideaux, D. (2003). ABC of learning and teaching in medicine. Curriculum design. *British Medical Journal, 326*(7383), 268-270.

Pulito, A. R., Donnelly, M. B., & Plymale, M. (2007). Factors in faculty evaluation of medical students' performance. *Medical education, 41*(7), 667-675.

Punch, K. (1998). *Introduction to social research: quantitative and qualitative approaches*. London: SAGE.

Quirk, M. & Babineau, R. (1982). Teaching interviewing skills to students in clinical years: a comparative analysis of three strategies. *Journal of Medical Education, 57*, 939-941.

Ram, P., Grol, R., Rethans, J., Schouten, B., van der Vleuten, C. & Kester, A. (1999). Assessment of general practitioners by video observation of communicative and medical performance in daily practice: issues of validity, reliability and feasibility. *Medical Education, 33*, 447-454.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4-13.

Ramsey, P. G., Wenrich, M. D., Carling, J. D., Inui, T. S., Larson, E. B., & LoGerfo, J. P. (1993). Use of peer ratings to evaluate physician performance. *The Journal of The American Medical Associtión, 269*(13), 1655–1660

Rasinski, K. A., Crocker, J., & Hastie, R. (1985). Another Look at Sex Stereotypes and Social Judgments: An Analysis of the Social Perceiver's Use of Subjective Probabilities. *Journal of Personality and Social Psychology, 49*(2), 317-326.

Read, S. J., Jones, D. K., & Miller, L. C. (1990). Traits as goal-based categories: The importance of goals in the coherence of dispositional categories. *Journal of Personality and Social Psychology, 58*(6), 1048-1061.

Reeder, G. D., Kumar, S., Hesson-McInnis, M., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology, 83*(4), 789-803.

Reeves, S., Albert, M., Kuper, A., & Hodges, B. D. (2008). Qualitative Research: Why Use Theories in Qualitative Research? *BMJ: British Medical Journal, 337*(7670), 631-634.

Regehr, G., MacRae, H., Reznick, R. K., & Szalay, D. (1998). Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Academic medicine, 73*(9), 993-997

Rethans, J. J., & van Boven, C. P. (1987). Simulated patients in general practice: a different look at the consultation. *British medical journal (Clinical research ed.), 294*(6575), 809-812

Reznick, R. K., Blackmore, D., Dauphinee, W., Rothman, A., & Smee, S. (1996). Large-scale high-stakes testing with an OSCE: report from the Medical Council of Canada. *Academic medicine, S71*, 19-21

Richardson, J. T. E. (2008). The attainment of ethnic minority students in UK higher education. *Studies in Higher Education, 33*(1), 33-48.

Richter Lagha, R. A., Boscardin, C. K., May, W., & Fung, C.-C. (2012). A comparison of two standard-setting approaches in high-stakes clinical performance assessment using generalizability theory. *Academic medicine : journal of the Association of American Medical Colleges, 87*(8), 1077-1082.

Roberts, C., Rothnie, I., Zoanetti, N., & Crossley, J. (2010). Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Medical education, 44*(7), 690.

Roberts, P., Priest, H., & Traynor, M. (2006). Reliability and validity in research. *Nursing standard (Royal College of Nursing (Great Britain) : 1987), 20*(44), 41

Roediger, H. L., Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382-395

Rosenberg, S., Nelson, C., & Vivekananthan, P. S. (1968). A multidimensional approach to the structure of personality impressions. *Journal of Personality and Social Psychology, 9*(4), 283-294

Ross, S., Poth, C. N., Donoff, M., Humphries, P., Steiner, I., Schipper, S., . . . Nichols, D. (2011). Competency-Based Achievement System Using formative feedback to teach and assess family medicine residents' skills. *Canadian Family Physician, 57*(9), e323.

Roter, D., Frankel, R., Hall, J. & Sluyter, D. (2006). The expression of emotion through non-verbal behaviour in medical visits. Mechanisms and outcomes. *Journal of General Internal Medicine, 21 (Suppl.1)*, 28-34.

Roter, D., Hall, J. & Katz, N. (1987). Relations between physicians' behaviours and analogue patients' satisfaction, recall, and impressions. *Medical Care, 25*, 437-451.

Rothbart, M., & John, O. (1993). Intergroup relations and stereotype change: A social-cognitive analysis and some longitudinal findings. In P. Sniderman, P. Tetlock & E. Carmines (Eds.), *Prejudice, politics and the American dilemma*. Stanford, CA: Stanford University Press.

Rushforth, H. E. (2007). Objective structured clinical examination (OSCE): Review of literature and implications for nursing education. *Nurse Education Today, 27*(5), 481-490.

Rutter, D. & Maguire, P. (1976). History-taking for medical students, II: Evaluation of a training programme. *Lancet, ii*, 558-560.

Sadler, D. R. (2010). Fidelity as a precondition for integrity in grading academic achievement. *Assessment & Evaluation in Higher Education, 35*(6), 727-743.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *instructional Science, 18*, 119-144.

Saettler, P. (1990). *The Evolution of American Educational Technology*. Englewood, CO: Libraries Unlimited.

Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in Black and White children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39*(4), 590-598.

Saldaña, J. (2013). *The coding manual for qualitative researchers* (Vol. 2nd). Los Angeles, Calif;London;: SAGE.

Sanson-Fisher, R. W., & Poole, A. D. (1980). Simulated patients and the assessment of medical students' interpersonal skills. *Medical education, 14*(4), 249-253

Sapsford, R. J., & Jupp, V. (2006). *Data collection and analysis* (Vol. 2nd). London: SAGE.

Scheff, T. (1967). *Mental Illness and Social Processes*. New York: Harper & Row.

Schleicher, D., Day, D., Mayes, B., & Riggio, R. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*(4), 735-746.

Schon, D. A. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.

Schneider, D., Hastorf, A., & Ellsworth, P. (1979). *Person perception* Reading, MA: Addison-Wesley.

Schreier, A. & Dub, B. (1981). Teaching interpersonal communication skills in pediatrics with the help of mothers. *South African Medical Journal, 59*, 865-866.

Schuh, L., London, Z., Neel, R., Brock, C., Kissela, B., & Schultz, L. (2009). Education research: Bias and poor interrater reliability in evaluating the neurology clinical skills examination. *Neurology, 73*(11), 904-908.

Schuwirth, L., & 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. (2012). Assessing competence. In B. Hodges & L. Lingard (Eds.), *The Question of Competence: Reconsidering Medical Education in the Twenty-First Century* (pp. 113-130). New York: Cornell University Press.

Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2006). Challenges for educationalists. *British Medical Journal, 333*(7567), 544-546

Schwandt, T. (1990). Paths to inquiry in the social sciences: scientific, constructivist, and critical theory methodologies. In E. Guba (Ed.), *The Paradigm Dialog* (pp. 258-276). Newbury Park, CA: Sage Publications.

Schwartzman, E., Hsu, D. I., Law, A. V. & Chung, E. P. (2011). Assessment of patient communication skills during OSCE: examining effectiveness of a training program in minimizing inter-grader variability. *Patient Education & Counseling, 83*(3), 472-477.

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the Latent Structure of Job Performance Ratings. *Journal of Applied Psychology, 85*(6), 956-970.

Secord, P. F., Bevan, W., & Katz, B. (1956). The Negro stereotype and perceptual accentuation. *The Journal of Abnormal and Social Psychology, 53*(1), 78-83.

Shanafelt, T. D., Bradley, K. A., Wipf, J. E., & Back, A. L. (2002). Burnout and self-reported patient care in an internal medicine residency program. *Annals of internal medicine, 136*(5), 358-367

Shanley, E. (2001). Misplaced confidence in a profession's ability to safeguard the public? *Nurse education today, 21*(2), 136-142

Shepard, L. A. (2000). The role of assessment in a learning culture. *educational Researcher, 29*, 4-14.

Sherman, J. W., Lee, A. Y., Bessenoff, G. R., & Frost, L. A. (1998). Stereotype efficiency considered: Encoding exibility under cognitive load. *Journal of Personality and Social Psychology, 75*, 589–606

Sherman, J. W., Stroessner, S. J., Loftus, S. T., & Deguzman, G. (1997). Stereotype suppression and recognition memory for stereotypical and non-stereotypical information. *Social Cognition, 15*(3), 205-215

Showers, C., & Cantor, N. (1985). Social Cognition: A Look at Motivated Strategies. *Annual review of psychology, 36*(1), 275-305.

Shumway, J. M., & Harden, R. M. (2003). AMEE Guide No. 25: the assessment of learning outcomes for the competent and reflective physician. *Medical Teacher, 25*(6), 569-584

Silber, C. G., Nasca, T. J., Paskin, D. L., Eiger, G., Robeson, M., & Veloski, J. J. (2004). Do global rating forms enable program directors to assess the ACGME competencies? *Academic medicine : journal of the Association of American Medical Colleges, 79*(6), 549-556.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131-142

Sloan, D. A., Donnelly, M. B., Schwartz, R. W., Felts, J. L., Blue, A. V. & Strodel, W. E. (1996). The use of objective structured clinical examination (OSCE) for evaluation and instruction in graduate medical education. *Journal of Surgical Research, 63*(1), 225-230.

Sloan, D. A., Donnelly, M. B., Schwartz, R. W. & Strodel, W. E. (1995). The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance. *Annals of Surgery, 222*(6), 735-742.

Smee, S. (2003). ABC Of Learning And Teaching In Medicine: Skill Based Assessment. *BMJ: British Medical Journal, 326*(7391), 703-706

Smith, E. R., & Semin, G. R. (2007). Situated Social Cognition. *Current Directions in Psychological Science, 16*(3), 132-135.

Smith, P., & Ragan, T. (1999). *Instructional Design* (2nd ed.). New York, NY: John Wiley & Sons.

Someren, M. W. v., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: a practical guide to modelling cognitive processes*. London: Academic Press.

Stangor, C. (1988). Stereotype Accessibility and Information Processing. *Personality and Social Psychology Bulletin, 14*(4), 694-708.

Stangor, C., Lynch, L., Duan, C., & Glass, B. (1992). Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology, 62*(2), 207-218.

Stewart, G. L., & Nandkeolyar, A. K. (2007). Exploring How Constraints Created by Other People Influence Intraindividual Variation in Objective Performance Measures. *Journal of Applied Psychology, 92*(4), 1149-1158

Stewart, M. A. (1995). Effective physician-patient communication and health outcomes: a review. *Canadian Medical Association Journal, 152*(9), 1423-1433.
Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute Identification by Relative Judgment. *Psychological Review, 112*(4), 881-911.
Stiggins, R. J. (1994). *Student-Centered Classroom Assessment*. New York: Merrill.
Stokes, J. (1974). The Clinical Examination - Assessment of Clinical Skills: Medical Education Booklet 2. Association for the Study of Medical Eucation, Dundee, UK.
Strack, F., Schwarz, N., Bless, H., Kübler, A., & Wänke, M. (1993). Awareness of the influence as a determinant of assimilation versus contrast. *European Journal of Social Psychology, 23*(1), 53-62.
Strauss, A. (1987). *Qualitative analysis for social scientists*. New York: Cambridge University Press.
Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research: techniques and procedures for developing grounded theory* (Vol. 2nd). Thousand Oaks: Sage Publications.
Strauss, A., & Corbin, J. M. (1990). *Basics of Qualitative Research: Grounded Theory Procedures and Techniques* (2nd ed.): SAGE Publications.
Streiner, D., & Norman, G. (2008). *Health measurement scales* (4th ed.). Oxford: Oxford University Press.
Sturman, M. C., Cheramie, R. A., & Cashen, L. H. (2005). The Impact of Job Complexity and Performance Measurement on the Temporal Consistency, Stability, and Test-Retest Reliability of Employee Job Performance Ratings. *Journal of Applied Psychology, 90*(2), 269-283
Surowiecki, J. (2005). *The Wisdom of Crowds*. New York, NY: Random House Digital.
Swanson, D. B. (1987). A measurement framework for performance based test. In I. Hart & R. Harden (Eds.), *Further Developments in Assessing Clinical Competence*. Montreal: Can-Heal.
Swanwick, T. (2010). *Understanding medical education: evidence, theory and practice*. Oxford: Wiley-Blackwell.
Tagiuri, R., Bruner, J. S., & Blake, R. R. (1958). On the relation between feelings and perception of feelings among members of small groups. In E. E. Maccoby & e. al (Eds.), *Readings in social psychology*. New York: Holt, Rinehart & Winston.
Tajfel, H. (1969). Cognitive Aspects of Prejudice. *Journal of Social Issues, 25*(4), 79-97.
Tamblyn, R., & Barrows, H. (1999). Data collection and interpersonal skills: The standardised patient encounter. In A. Tekian, C. McGuire & W. McGaghie (Eds.).
Tanner, C. (2008). Context Effects in Environmental Judgments: Assimilation and Contrast Effects in Separate and Joint Evaluation Modes. *Journal of Applied Social Psychology, 38*(11), 2759-2786.
Tavares, W., & Eva, K. W. (2013). Exploring the impact of mental workload on rater-based assessments. *Advances in Health Sciences Education, 18*(2), 291-303.
Taylor, C. (1994). Assessment for measurement or standards: the peril and promise of large-scale assessment reform. *American Educational Research Journal, 31*, 231-262.
Tekian, A. (1999). Assessing communication, technical, and affective responses: Can they relate like a professional? In A. Tekian, C. McGuire & W. McGaghie (Eds.), *Innovative simulations for assessing clinical competence* (pp. 105-112). Chicago: Department of Medical Education, University of Illinois at Chicago.
ten Cate, O., & Scheele, F. (2007). Competency-based postgraduate training: can we bridge the gap between theory and clinical practice? *Academic medicine : journal of the Association of American Medical Colleges, 82*(6), 542
ten Cate, O., Snell, L., & Carraccio, C. (2010). Medical competence: the interplay between individual ability and the health care environment. *Medical Teacher, 32*(8), 669–675
Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45*(1), 74-83.

Tetlock, P. E., & Kim, J. I. (1987). Accountability and Judgment Processes in a Personality Prediction Task. *Journal of Personality and Social Psychology, 52*(4), 700-709.

Thammasitboon, S., Mariscalco, M., Yudkowsky, R., Hetland, M., Noronha, P., & Mrtek, R. (2008). Exploring individual opinions of potential evaluators in a 360-degree assessment: Four distinct viewpoints of a competent resident. *Teaching and learning in medicine, 20*(4), 314-322.

Todd, A. R., Bodenhausen, G. V., Richeson, J. A., & Galinsky, A. D. (2011). Perspective taking combats automatic expressions of racial bias. *Journal of Personality and Social Psychology, 100*(6), 1027-1042.

Torre, D., Daley, B., Sebastian, J., & Elmicki, D. (2006). Overview of current learning theories for medical educators. *American Journal of Medicine, 119*, 903-907.

Townsend, A. H., McLlvenny, S., Miller, C. J., & Dunn, E. V. (2001). The use of an objective structured clinical examination (OSCE) for formative and summative assessment in a general practice clinical attachment and its relationship to final medical school examination performance. *Medical education, 35*(9), 841-846

Turner, J. L. & Dankoski, M. E. (2008). Objective structured clinical exams: a critical review. *Family Medicine, 40*(8), 574-578.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science, 185*(4157), 1124-1131.

Tweed, M., & Ingham, C. (2010). Observed consultation: confidence and accuracy of assessors. *Advances in Health Sciences Education, 15*(1), 31-43.

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago, IL: University of Chicago Press.

Tziner, A., Murphy, K. R., & Cleveland, J. N. (2005). Contextual and Rater Factors Affecting Rating Behavior. *Group & Organization Management, 30*(1), 89-98

Uggerslev, K. L., & Sulsky, L. M. (2008). Using Frame-of-Reference Training to Understand the Implications of Rater Idiosyncrasy for Rating Accuracy. *Journal of Applied Psychology, 93*(3), 711-719.

Ugwuegbu, D. C. E. (1979). Racial and evidential factors in juror attribution of legal responsibility. *Journal of Experimental Social Psychology, 15*(2), 133-146.

Uleman, J. S., Adil Saribay, S., & Gonzalez, C. M. (2008). Spontaneous inferences, implicit impressions, and implicit theories. *Annual review of psychology, 59*(1), 329-360.

van Barneveld, C. (2005). The dependability of medical students' performance ratings as documented on in-training evaluations. *Academic Medicine, 80*(3), 309-312.

van den Bergh, L., Denessen, E., Hornstra, L., Voeten, M., & Holland, R. W. (2010). The Implicit Prejudiced Attitudes of Teachers: Relations to Teacher Expectations and the Ethnic Achievement Gap. *American Educational Research Journal, 47*(2), 497-527.

van der Vleuten, C. (1996). The Assessment of Professional Competence: Developments, Research and Practical Implications. *Advances in Health Sciences Education, 1*, 41-67.

van der Vleuten, C. P., Norman, G. R., & De Graaff, E. (1991). Pitfalls in the pursuit of objectivity: issues of reliability. *Medical education, 25*(2), 110-118

van der Vleuten, C. & Schuwirth, L. (2005). Assessing professional competence: From methods to programmes. *Medical Education, 39*(3), 309-317.

van der Vleuten, C. P. M., & Swanson, D. B. (1990). Assessment of clinical skills with standardized patients: State of the art. *Teaching and learning in medicine, 2*(2), 58-76

van der Vleuten, C. P., van Luyk, S. J., van Ballegooijen, A. M., & Swanson, D. B. (1989). Training and experience of examiners. *Medical education, 23*(3), 290-296

van der Vleuten, C. P. M. v. d., Durning, S. J., Boulet, J. R., Dorrance, K., Schuwirth, L., & Artino, A. R. (2012). The impact of selected contextual factors on experts' clinical reasoning performance (does context impact clinical reasoning performance in experts?). *Advances in Health Sciences Education, 17*(1), 65-79.

van der Vleuten, C. P. M. v. d., Kramer, A. W. M., Mokkink, H. G. A., Grol, R. P. T. M., Pelgrim, E. A. M., & Elsen, L. v. d. (2011). In-training assessment using direct observation of single-patient encounters: a literature review. *Advances in Health Sciences Education, 16*(1), 131-142.

van Merriënboer, J. J. G., & Sweller, J. (2010). Cognitive load theory in health professional education: design principles and strategies. *Medical education, 44*(1), 85.

van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes (Knowledge-Based Systems* (1st ed.). London: Academic Press.

Van Thiel, J., van der Vleuten, C. P. M. & Kraan, H. (1992). Assessment of medical interviewing skills: Generalizability of scores using successive MAAS-versions. In R. M. Harden, I. R. Hart & H. Mulholland (Eds.), *Approaches to assessment of clinical competence-Part II*. Norwich: Page Brothers.

Veldhuijzen, W., Ram, P., van der Weijden, T., Niemantsverdriet, S., & van der Vleuten, C. (2007). Characteristics of communication guidlines that facilitate or impede guideline use: A focus group study. *BMC Family Practice, 8*, 31.

Viswesvaran, O., Ones, D., & Schmidt, F. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*(5), 557-574.

Vu, N. V., & Barrows, H. S. (1994). Use of Standardized Patients in Clinical Assessments: Recent Developments and Measurement Findings. *Educational Researcher, 23*(3), 23-30

Vukanovic-Criley, J. M., Criley, S., Warde, C. M., Broker, J. R., Guevara-Mathews, L., Churchill, W. H., . . . Criley, J. M. (2006). Competency in cardiac examination skills in medical students, trainees, physicians, and faculty: a multicentre study. *Arch Intern Med, 166*(6), 610–616.

Wass, V., Bowden, R. & Jackson, N. (2007). The principles of assessment design. In N. Jackson, A. Jamieson & A. Khan (Eds.), *ASSESSMENT IN MEDICAL EDUCATION AND TRAINING*. Oxford: Radcliffe Publishing Ltd.

Wass, V., & van der Vleuten, C. (2004). The long case: the metric of medical education. *Medical education, 38*(11), 1176-1180

Wass, V., van der Vleuten, C., Shatzer, J. & Jones, R. (2001). Assessment of clinical competence. *Lancet, 357*(9260), 945-949.

Wass, V., Vleuten, C., Shatzer, J., & Jones, R. (2001). Medical education quarter assessment of clinical competence. *Lancet, 357*, 945-949.

Weaver, K., & Olsen, J. (2006). Understanding paradigms used for nursing research. *journal of Advanced Nursing, 53*(4), 459-469.

Webster-Wright, A. (2009). Reframing Professional Development through Understanding Authentic Professional Learning. *Review of Educational Research, 79*(2), 702-739.

Wedell, D. H., Santoyo, E. M., & Pettibone, J. C. (2005). The Thick and the Thin of It: Contextual Effects in Body Perception. *Basic and Applied Social Psychology, 27*(3), 213-228.

Wegner, D. M. (1994). Ironic Processes of Mental Control. *Psychological Review, 101*(1), 34-52

Weiner, B. (1995). Inferences of responsibility and social motivation. In M. Zanna (Ed.), *Advances in Experimental Social Psychology, vol 27* (pp. 1-47). San Diego, CA: Academic Press.

Weller, J., Jolly, B., Misur, M., Merry, A., Jones, A., & Crossley, J. (2009). Mini-clinical evaluation exercise in anaesthesia training. *British Journal of Anaesthesia, 102*(5), 633-641.

Wessel, J., Williams, R., Finch, E., & Gémus, M. (2003). Reliability and validity of an objective structured clinical examination for physical therapy students. *Journal of Allied Health, 32*(4), 266-269

Whelan, G. (1999). Educational commission for foreign medical graduates: Clinical skills assessment prototype. *Medical Teacher, 21*, 156–160

Wiley, D. E. & Haertel, E. H. (1996). Extended assessment tasks: purposes, definitions, scoring and accuracy. In M. B. Kane & R. Mitchell (Eds.), *Implementing performance assessment: promises, problems and challenges*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wiliam, D. (2000). Integrating summative and formative functions of assessment. *Assessment in Education, 9*(58), 10.
Wiliam, D. (2003a). *Assessment and learning*. Paper presented at the ATL Summer Conference; London, UK.
Wiliam, D. (2003b). *National assessment: how to make it better*. Paper presented at the RSA; London, UK.
Wiliam, D. (2001). Reliability, validity, and all that jazz. *Education 3-13, 29*(3), 17-21
Wiliam, D. (2008). *When is assessment learning-oriented?* Paper presented at the 4th Biennial EARLI/Northumbria Assessment Conference; Potsdam, Germany.
Wiliam, D. (2011). *Formative assessment: definitions and relationships*. Paper presented at the Annual meeting of the American Educational Research Association; New Orleans, US.
Wilkinson, J., Crossley, J., Wragg, A., Mills, P., Cowan, G., & Wade, W. (2008). Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical education, 42*(4), 364-373.
Williams, R. G., Klamen, D. A. & McGaghie, W. C. (2003). Cognitive, social and environmental sources of bias in clinical performance ratings. *Teaching & Learning in Medicine, 15*(4), 270-292.
Williams, S., Weinman, J. & Dale, J. (1998). Doctor-patient communication and patient satisfaction: A review. *Fam Pract, 15*, 480-492.
Willig, C., & Rogers, W. S. (2008). *The Sage handbook of qualitative research in psychology*. London: SAGE.
Wilmot, A. (2005). Designing sampling strategies for qualitative social research: with particular reference to the Office for National Statistics' Qualitative Respondent Register. *Survey Methodology Bulletin, 56*
Wilson, T. D. (1994). The proper protocol: Validity and completeness of verbal reports. *Psychological Science, 5*(5), 249-252.
Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology, 47*(2), 237-252
Wittenbrink, B., Gist, P. L., & Hilton, J. L. (1997). Structural Properties of Stereotypic Knowledge and Their Influences on the Construal of Social Situations. *Journal of Personality and Social Psychology, 72*(3), 526-543.
Wittenbrink, B., Park, B., & Judd, C. (1998). The role of stereotypic knowledge in the construal of person models. In C. Sedikides, J. Schopler & C. Insko (Eds.), *Intergroup cognition and intergroup behavior*. Mahawa, NJ: Lawrence Erlbaum Associates.
Woehr, D. J. & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*(3), 189-205.
Wojciszke, B. (2005). Affective Concomitants of Information on Morality and Competence. *European Psychologist, 10*(1), 60-70
Wolf, F. (2004). Methodological Quality, Evidence, and Research in Medical Education (RIME). *Academic medicine, 79*(10), S68-S69
Wood, L., Hassell, A., Whitehouse, A., Bullock, A., & Wall, D. (2006). A literature review of multi-source feedback systems within and without health services, leading to 10 tips for their successful design *Medical Teacher, 28*, 185-191.
Wood, T. J. (2013). Mental workload as a tool for understanding dual processes in rater-based assessments. *Advances in Health Sciences Education, 18*(3), 523-525.
Wood, T. J. (2014). Exploring the role of first impressions in rater-based assessments. *Advances in Health Sciences Education, 19*(3), 409-427.
Woolf, K., Potts, H. W. W., & McManus, I. C. (2011). Ethnicity and academic performance in UK trained doctors and medical students: systematic review and meta-analysis. *BMJ: British Medical Journal, 342*(7797), 584-584.
World Health Organization. (1993). Retrieved December 13, 2014, from

Ybarra, O., Chan, E., Park, D., Burnstein, E., Monin, B., & Stanik, C. (2008). Life's recurring challenges and the fundamental dimensions: An integration and its implications for cultural differences and similarities. *European Journal of Social Psychology, 38*(7), 1083-1092

Yeates, P., Moreau, M., & Eva, K. (2015). Are Examiners' Judgments in OSCE-Style Assessments Influenced by Contrast Effects? *Academic medicine : journal of the Association of American Medical Colleges, 90*(7), 975.

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2012). Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. *JAMA, 308*(21), 2226–2232

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013a). 'You're certainly relatively competent': assessor bias due to recent experiences. *Medical education, 47*(9), 910-922.

Yeates, P., O'Neill, P., Mann, K., & Eva, K. (2013b). Seeing the same thing differently: Mechanisms that contribute to assessor differences in directly-observed performance assessments. *Advances in Health Sciences Education, 18*(3), 325-341.

Yzerbyt, V. Y., Schadron, G., Leyens, J.-P., & Rocher, S. (1994). Social Judgeability: The Impact of Meta-Informational Cues on the Use of Stereotypes. *Journal of Personality and Social Psychology, 66*(1), 48-55.

Zubin, A., Carol, O. B., John, P., & Lila, Q. M. (2003). Development and Validation Processes for an Objective Structured Clinical Examination (OSCE) for Entry-to-Practice Certification in Pharmacy: The Canadian Experience. *American Journal of Pharmaceutical Education, 67*(3), 1-8

Appendix 1

How the OSCE works in Leeds*

OSCEs are utilised as part of a wider programme of assessment. Different assessment methods, called a 'test battery' approach, can be used and the OSCE is considered an essential examination in this test battery in the assessment of clinical performance in a simulated experience. The utility and usage of the OSCE is evidence based, and one that Leeds continues to contribute to.

The OSCE is used in Leeds in Years 2,3,4 and 5. Year 2 is merely for providing feedback to learners and introducing them to the test format and construct so they get familiar with before they get examined in summative and high stakes examinations. The OSCE is also used as a 'sandbox' in which variant approaches to gaining examiner feedback for learners are tested. Years 3,4 and 5 are all high stakes examinations conducted at the end of each academic year in order to help determine student progression and graduation. Year 3 is a 'traditional' large scale OSCE with the opportunity of taking the examination again. Years 4 and 5 are sequential testing formats in which a shorter screening test is taken by all candidates. The passing threshold in these high stakes examinations is higher and weaker candidates end up taking a longer/extended OSCE.

Blueprinting is carefully done for each OSCE (difficulty rises with candidate ability and year of study). Different skills are integrated in each OSCE station to enable examining more than just one skill. This integration of skills rises with later years of study. For instance, candidates in year 3 are asked to summarise and attempt to reach a diagnosis,

whereas year 5 is highly integrated with multifaceted constructs that examine variant skills and traits across the OSCE. The length of each station is decided based on the year of study and task.

The OSCE stations in Leeds are ensured to be highly authentic. Such stations could involve either Simulated or Real patients, and many are authored jointly between patients and carer group and clinicians. Each station uses some form of detailed scoring system (typically key features checklist) coupled with an assessor global grade. My research investigated how this global grade can be influenced by several and multifaceted factors. Patients are also typically asked to give a global grade on performance.

Examiners are recruited from local clinical teaching faculty. In addition to medical doctors, pharmacists, nurses and skills staff can examine. They are all trained with regard to the principles of OSCEs, examiner behaviours, scoring and standards. This standard training programme is customised to support those doing Year 3 or the sequential formats in Years 4 and 5. It is possible for some examiner, here in Leeds, to examine across all 3 years, while others may only examine one year of the programme because of clinical discipline or seniority. However, most examiners will see and teach students across more than one year of the programme. Therefore, extensive examiner support 'in station' is important to assist them with expected standards as well as on the day/bespoke examiner training and briefing before the beginning of the exam.

* Information obtained from my first supervisor, Professor Richard Fuller.

Appendix 2

Objectivity and Standardisation

* Objectivity here, the opposite of subjectivity, refers to judgment that is based on observable performance and not influenced by emotions or personal prejudices. All candidates get the same exam and are compared against a certain predefined criteria.

* Standardisation in the OSCE refers to the process of using a standard or a reference point against which performance can be assessed in a simulated environment.

The traditional approach to clinical assessment, the 'long case', was described as unfair due to examiner bias which makes it less reliable. Therefore, Harden sought to provide a fairer exam format for candidates (Harden et al., 2015). The previous two principles used in the OSCE can be linked with an aim to reduce examiner variance/inconsistency. This could ultimately increase reliability and fairness.

Appendix 3

Interview schedule (Examples of questions)

Theme	Question
The assessor as trainable	**Has your own experience as a patient or one of your relatives informed your decision as an assessor?**
The assessor as fallible	**How do you find the difference between video and face to face observations?** **How do you think male students are different from female students in terms of consultation skills, for example?** **How about your experience in assessing students from different cultures or religions?**
The assessor as idiosyncratic	**Would you give the student any feedback? What is it about?** **Could you please write the student character?** **Could you justify your decision?**

Appendix 4

Grade Descriptors

Clear Fail:

. Little idea of how to approach the station.

. Disorganized approach, no evidence of planning – tends to random actions, process and actions.

. Unable to synthesize findings, or reach a diagnosis/plan.

. Struggle/no response to questions about applied knowledge.

Borderline:

. Able to commence station, but often uncertain, and struggles to proceed to completion.

. Some organization of approach, but 'formulaic' with no flexibility (e.g. 'lists' of questions for patients)

. No evidence of reasoning/discrimination when answering questions in the station (e.g. unstructured 'lists')

Clear Pass:

. Systematic overall approach to station/task.

. Demonstrates sufficient organization to permit completion of task with some evidence of flexibility of approach.

. Able to summarize (e.g. present history/explain) and manage additional questioning with evidence of reasoning.

Very Good Pass:

. Clearly professional approach to station. Good levels of organization with clear evidence of flexibility.

. Clearly able to synthesize findings, or reach a diagnosis/plan.

. Clear evidence of planning, ability to summarize and manage questioning.

Excellent:

. Overall superior approach – excellent organizational skills, and fluent management of task in hand.

. Flexible, adaptive approach to changing circumstances within a station – e.g. reacting to patients, emergency situations.

. High levels of professionalism and clinical reasoning – applies knowledge critically when questioned.

Appendix 5

Coding (Example)

Student- Intro

Student- Elicit concerns

‘‘She made a clear introduction, she asked the patient what she wanted

Student- Listening

Student- Respond to Q

to know and listened to the answers, answered those questions and

Student- Thoroughness

then was able to add some more information so I thought that was all

Assessor- dislike

very…all very positive. The only thing I didn’t like at all in the whole

Assessor- influence on marks

station the thing that would have made it the bottom of a good pass

instead of it being an excellent is I couldn’t quite understand what she

Student- Distractor (mobile)

was doing with her mobile phone and in the middle of it she seemed to

Student- Distractors (fiddling)

Student- Distractor (watch) + Dedication

do a lot of hand fiddling and also was very overtly looking at her watch

Student- Respect

which is fine if you explain to the patient what…pardon me, why you’re

doing it. If you say, “I’ve been told I can only spend 10 minutes with you

do you mind if I just keep an eye on the time?’’ ”

Appendix 6

Ethical approval

Faculty of Medicine and Health

Research Office
University of Leeds
Worsley Building
Clarendon Way
Leeds LS2 9NL
United Kingdom

✆ +44 (0) 113 343

UNIVERSITY OF LEEDS

11 March 2013

Mr Sami Alnasser
PhDStudent
Medicine / Leeds Institute of Medical Education (LIME)
Worsley Building
University of Leeds

Dear Sami

Ref no: EDREC/12/017

Title: Influence of non-verbal communications on assessors' global marking in Objective Structured Clinical Examination

I am pleased to inform you that the above research application has been reviewed by the Medicine and Dentistry Educational Research Ethics Committee (EdREC) and I can confirm a favourable ethical opinion based on the documentation received at date of this letter.

However, please note that the Committee has recommended that separate consent forms are used for the focus group and the interviews.

Document	*Version*	*Date*
Ethical Review Form	1	28.01.13
Proposal	1	28.01.13
Participant Information Sheet	1	28.01.13
Consent Form	1	28.01.13
Response to Reviewer 2 comments	1	27.02.13
Consent Form	1	27.02.13

Please notify the committee if you intend to make any amendments to the original research as submitted at date of this approval. This includes recruitment methodology and all changes must be ethically approved prior to implementation. Please contact the Faculty Research Ethics and Governance Administrator for further information (fmhuniethics@leeds.ac.uk)

Ethical approval does not infer you have the right of access to any member of staff or student or documents and the premises of the University of Leeds. Nor does it imply any right of access to the premises of any other organisation, including clinical areas. The committee takes no responsibility for you gaining access to staff, students and/or premises prior to, during or following your research activities.

Please note: You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited.

It is our policy to remind everyone that it is your responsibility to comply with Health and Safety, Data Protection and any other legal and/or professional guidelines there may be.

I wish you every success with the project.

Yours sincerely

Dr John Sandars, Chair, EdREC

Appendix 7

Information Sheet

Invitation

Dear Sir/Madam,

You are being invited to take part in a research project. Before you decide it is important for you understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask me if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

The title of the research project

How non-verbal behaviours influence assessors' global marking when they observe and assess undergraduate medical students using objective structured clinical examinations.

The purpose of the study

I am a PhD student doing my research in Medical Education. Particularly, my area of interest is assessment in medical education and my research question is trying to find out what and how non-verbal behaviours influence assessors' global marking in the OSCE. The length of the degree is between three and four years. I am now in my second year collecting the required data for my research.

Why have you been chosen?

In order to answer the research question, and as mentioned earlier, I am looking at what and how non-verbal behaviours influence assessors' global marking in the OSCE. Therefore, I am required to interview National Health Service staff who act as university assessors.

Do you have to take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time. You do not have to give a reason.

What will happen to you if you take part?

Qualitative research methods will be utilised to gather a thorough understanding of human behaviours and reasons that govern such behaviours. You will be interviewed only one time (around 45-60 min long) to discuss some points that can help in answering the research question. Open ended questions related to the topic will be asked. Two short video clips of a

student communicating with a simulated patient will be shown in order to further facilitate the discussion and collection of data.

What are the possible benefits of taking part?

Reflection is defined by Reid (1993) as "a process of reviewing an experience of practice in order to describe, analyse, evaluate and so inform learning about practice''. Impact on own assessment behaviour is a possible benefit of taking part in this project as reflection helps in improving and enhancing the way individuals teach and assess. This is mainly because of the possible positive change in behaviour in future after evaluating an own experience.

Will your taking part in this project be kept confidential?

All the information that we collect about you during the course of the research will be kept strictly confidential. You will not be able to be identified in any reports or publications. We will always use numbers to refer to the participants and not by their names.

Will you be recorded?

There will be audio recording. It will be used only for analysis. No other use will be made of without your written permission, and no one outside the project will be allowed access to the original recordings.

Thank you very much for taking your time to read through this information sheet

For further information

umssaln@leeds.ac.uk

Appendix 8

Consent to take part in a PhD project about:

"How non-verbal behaviours influence assessors' global marking when they observe and assess undergraduate medical students using objective structured clinical examinations

I agree that I have read and understood the information sheet explaining the above research project and I have had the opportunity to ask questions about the project.

I understand that my participation is voluntary and that I am free to withdraw at any time without giving any reason and without there being any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline.

I give permission for members of the research team to have access to my anonymous responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research. I understand that my responses will be kept strictly confidential.

I agree for the data collected from me to be audibly recorded and used in relevant future research.

I agree to take part in the above research project and will inform the lead researcher should my contact details change.

Name of participant Date:

Participant's signature

Name of lead researcher Date:

Signature

Participant's allocated number:

Appendix 9

Information sheet and consent form (students)

Influence of non-verbal behaviours on assessors' global marking in the OSCE

What is required of you?

This project is about non-verbal behaviours that could influence OSCE assessors' global judgements. A video clip of a medical student communicating with a simulated patient is required to further facilitate discussion and data collection.

Consent

I agree that I have read and understood the above, and I have had the opportunity to ask questions about the project or what I need to do.

I understand that my participation is voluntary.

I give permission for members of the research team to videotape me while I am communicating with a simulated patient.

I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.

I agree to take part in the above research project.

Name Date

Signature

Appendix 10

Information sheet and consent form (simulated patient)

Influence of non-verbal behaviours on assessors' global marking in the OSCE

What is required from you?

This project is about non-verbal behaviours that could influence OSCE assessors' global judgements. A video clip of a medical student communicating with a simulated patient is required to further facilitate discussion and data collection.

Consent

I agree that I have read and understood the above, and I have had the opportunity to ask questions about the project or what I need to do.

I understand that my participation is voluntary.

I give permission for members of the research team to videotape me while I am communicating with a medical student.

I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research.

I agree to take part in the above research project.

Name Date

Signature

Appendix 11

Poster

(Presented in Glasgow - AMEE 2015)

UNIVERSITY OF LEEDS

It does not need to be voiced to be counted!

How non-verbal behaviour influences assessors' global marking when examining undergraduate medical students using objective structured clinical examinations.

Sami Alnasser, Richard Fuller, Trudie Roberts

Background & objective

Objective Structured Clinical Examinations (OSCEs) are one of the most common performance assessment tools, but can be subject to a variety of potential threats to their reliability. Whilst differences in assessor judgments have been labelled as 'error variance', recent workplace assessment based research has explored different perspectives of assessors (and their decisions) through constructivist lenses, identifying them as 'trainable, fallible or meaningfully idiosyncratic' (Gingerich et al. 2014).

Perspectives from psychosocial research have explored factors influencing this idiosyncrasy, but less attention has been paid to non-verbal behaviours of candidates, assessors and patients that could influence assessors judgments during OSCEs. We investigated how non-verbal behaviour influences assessors' global marking when examining undergraduate medical students using objective structured clinical examinations.

Summary of work

18 OSCE assessors participated in in-depth interviews. Each scored 2 videos of students consulting with a simulated patient, and made judgments on each performance using a standard checklist and written feedback. A retrospective think aloud methodology was used as a stimulus to explore factors in the students' performances. Interview transcripts were coded and a grounded theory approach used to develop a framework to interpret results (Glaser & Strauss, 1967).

Results

Thematic analysis revealed a rich framework where the interaction and non-verbal behaviours of assessors, patients and candidates all contributed to global ratings. Assessors' identification and response to candidate behaviours was complex and individual. Subthemes included, for example, the importance on 'body language' and the impact of assessor fatigue, coupled with the use of pre-determined stereotypes.

Conclusion

The nonverbal behaviours of the three 'characters' in the OSCE (student, patient and assessor) make significant contributions to global ratings. This has importance in station and scoring format design, assessor training and the ongoing research into the assessor decision making in high stakes performance tests.

Take home message

It does not need to be voiced to be counted.

Non-verbal behaviours within an OSCE station have a significant impact on assessor judgements, and contribute to the multiple factors that reduce inter-rater reliability.

References

1-Gingerich et al. 2014, Seeing the black box differently: assessor cognition from three research perspectives. Medical Education, 48: 1055-1068

2- Denny et al. 2013, MRCGP CSA: are the examiners biased, favouring their own sex, ethnicity, and degree source? British Journal of General Practice. e718-e725

3- Govaerts et al. 2013, Workplace-based assessment: raters' performance theories and constructs. Adv in Health Sci Educ, 18: 375-396

4- Kenny DA, 1994, Interpersonal Perception: A Social Relations Analysis. New York, NY: Guilford Press.

5- Glaser and Strauss, 1967, The discovery of grounded theory: strategies for qualitative research. Chicago: Aldine