

The Assessment of Communication Skills During OSCEs: Development and Trialled Implementation of a New Standardised Model Using the MAAS-Global Instrument

Winnie Setyonugroho BMed MEng

A thesis submitted for the degree of Doctor of Philosophy

Supervisors of research:

Dr Thomas JB Kropmans PT PPT MSc PhD
Dr Kieran M Kennedy MB MICGP MSc MHSc MMedSc

Medical Informatics & Education, School of Medicine
College of Medicine, Nursing and Health Sciences
National University of Ireland, Galway

January 2016

Contents

List of Tables.....	iv
List of Figures	v
Declaration.....	vi
Acknowledgements	vii
Abstract.....	ix
Chapter 1. General Introduction.....	1
1.1. Introduction.....	1
1.2. Rationale.....	1
1.2.1. Communication Skills	1
1.2.2. OSCE	2
1.2.3. The CS Assessment.....	4
1.2.4. CS Measurement Instrument.....	5
1.2.5. Calibration of Measurement Instrument.....	7
1.2.6. The Effect of L1 and L2 in Students Performance.....	7
1.3. General Aim, Research Questions, and Outline of the Thesis	9
1.4. References.....	13
Chapter 2. Reliability and Validity of OSCE Checklists Used to Assess the Communication Skills of Undergraduate Medical Students: a Systematic Review	16
2.1. Abstract	16
2.2. Background.....	17
2.3. Method.....	19
2.4. Results	24
2.4.1. Search Results	24
2.4.2. Content Analysis.....	25
2.4.3. Standard Setting.....	34
2.4.4. Reviewer Agreement	34
2.5. Discussion and Conclusion	36
2.5.1. Discussion.....	36
2.5.2. Conclusion	40
2.6. Practice Implications	41
2.7. References for the Study	41
2.8. References for the Review	44
Chapter 3. Calibration of Communication Skills Items in OSCE Checklists According to the MAAS-Global.....	47
3.1. Abstract	47
3.2. Background.....	48
3.3. Methods	49
3.3.1. Context of the Study	49
3.3.2. Description of the OSCE	49
3.3.3. Calibration Checklists	51

3.3.4.	Choice of Statistical Approach	52
3.3.5.	Procedure	53
3.3.6.	Analysis.....	54
3.4.	Results	54
3.4.1.	Station Checklists	54
3.4.2.	Reliability Analysis.....	55
3.4.3.	G-Kappa Results for Discipline Checklists	56
3.4.4.	Communication Skills within Station Checklists	56
3.5.	Discussion and Conclusion	57
3.5.1.	Discussion.....	57
3.5.2.	Limitation of the Study.....	63
3.5.3.	Conclusion	64
3.6.	Practice Implications	64
3.7.	References.....	65
Chapter 4. True Communication Skills Assessment in Interdiscipline OSCE Stations: Standard Setting Using The MAAS-Global and EduG.....		68
4.1.	Abstract	68
4.2.	Background.....	69
4.3.	Methods	71
4.3.1.	Context of the Study	71
4.3.2.	Research Subjects	71
4.3.3.	Measurement Instrument.....	72
4.3.4.	Conversion Method.....	72
4.3.5.	Statistical Analysis	73
4.4.	Results	73
4.4.1.	Station Characteristics	73
4.4.2.	Scores	76
4.4.3.	Reliability Analysis.....	78
4.5.	Discussion	79
4.6.	Conclusion	83
4.7.	References.....	83
Chapter 5. The Effect of Language on the Assessment of Communication Skills Using a Standardised Measurement Instrument in OSCEs.....		85
5.1.	Abstract	85
5.2.	Background.....	86
5.3.	Method	89
5.3.1.	Context of the Study	89
5.3.2.	Research Subjects	90
5.3.3.	Rubrics Calibration	91
5.3.4.	The Standard	91
5.3.5.	Statistical Analysis	91
5.4.	Results	92
5.4.1.	The OSCEs.....	92

5.4.2.	Rubrics Calibration	94
5.4.3.	MAAS-Global Score	95
5.5.	Discussion	101
5.6.	Limitations of the Study	106
5.7.	Conclusion	107
5.8.	References	107
Chapter 6.	General Conclusion	111
6.1.	Preface	111
6.2.	General Discussion	112
6.2.1.	A standard of Communication Skills (CS) Measurement Instrument 112	
6.2.2.	Calibration Method and Procedures.....	113
6.2.3.	Standardised Communication Skills.....	115
6.2.4.	MAAS-Global Characteristics and the Second Language (L2).....	118
6.2.5.	The Reliability of the OSCEs	121
6.3.	Generalisation	122
6.4.	Implications for Further Research.....	122
6.5.	Recommendation for CS Assessment in OSCEs.....	125
6.6.	Conclusion	125
6.7.	Dissemination and Other Achievements During the Project	126
6.8.	References	129
Appendix 1.	MAAS-Global check-list calibration manual.....	132
Appendix 2.	Ethical approval.....	144

List of Tables

Table 2.1: Steps conducted in the initial narrative review to retrieve the appropriate literature for the systematic critical appraisal of the literature.....	20
Table 2.2: Details of papers included in the systematic review and an overview of the communication skills domains reported in each.....	27
Table 2.3: Communication skills domains agreement between 3 reviewers ICC= Intra Class Correlation coefficient Full agreement between reviewers (ICC = 1) means 100% agreement on items assessed. No agreement (ICC = 0) reviewers don't agree at all on items assessed.	33
Table 3.1: Summary of OSCE's circuits, stations, and checklist's items in each disciplines in 3 academic terms (2010/2011, 2011/2012, and 2012/2013).....	50
Table 3.2: Summary of estimated variance component (G-Study), G-Kappa coefficient, and D Study (optimization) Analysis.	58
Table 3.3: MAAS-Global sections and items of stations' checklists in each discipline.	59
Table 4.1: Summary of OSCE score Mean, MAAS-Global score Mean, Communication skills, MAAS-Global proportion, section of MAAS-Global in percentage, and MAAS-Global items.....	75
Table 4.2: Analysis of variance for both OSCE score and MAAS-Global score (compare the effect of disciplines and circuits which nested within disciplines).	77
Table 4.3: Summary of Generalizability Coefficient and Decision Study (with 10 and 15 stations).....	78
Table 5.1: Summary of L1 and L2 students' country of origin.	92
Table 5.2: Summary of MAAS-Global proportion, sections and items of stations' checklists in each discipline.	96
Table 5.3: Summary of OSCE settings, Generalizability Coefficient and Decision Study (with 10 and 15 stations), and MAAS-Global score (L1 and L2 students). .	99

List of Figures

Figure 1.1: Miller's Pyramid.	3
Figure 2.1: The different stage of the systematic review.	24
Figure 2.2: Comparison between resolved disagreement and initial agreement amongst reviewers. * V / R = validity / reliability; CS Domains = Communication Skills domains.....	35
Figure 3.1: Illustration on how raters calibrate station's checklists and then transferred into spreadsheet.	53

Declaration

This work is submitted to fulfill the requirements of the degree of Doctor of Philosophy at the National University of Ireland, Galway. No part of this thesis has been previously submitted at this or at any other university. Apart from due acknowledgements, it is entirely my own work.

Signed: _____
Winnie Setyonugroho

Date: _____

Acknowledgements

This completion of this dissertation was definitely one of the hardest tasks of my life so far. I would never been able to finish this dissertation without the dedicated guidance of my supervisors. I would like to express my sincere gratitude to Dr. Thomas JB Kropmans for his continuous support of my research and PhD study. His guidance helped me since the first day I entered the NUI Galway. During the first day of my arrival in Galway, a beautiful rainbow appeared at the horizon, wishing me “good luck” we are now 3.5 years down the route of my “pot of gold at the end of the rainbow”. I would also like to thank my co-supervisor, Dr. Kieran M. Kennedy MB MICGP MSc MHSc MMedSc, for his insightful thoughts these past several years.

Besides my supervisor, I am grateful for the invaluable work of all co-authors of my papers published and submitted, who provided me with guiding directions, insightful thoughts and positive feedback on the English spelling and grammar. Without their precious support to create high quality papers, it would not have been possible to complete this thesis. Special thanks to Dr. Jan van Dalen, the creator of MAAS-Global, who was willing to participate and provide guidance in writing the background theory of two of my papers. I would like also to thank Dr. Maureen Kelly PhD, who helped develop my background knowledge in language and student performance in my fourth paper.

My sincere thanks to Prof. Jean Cardinet, The Institute for Educational Research and Documentation (IRD), who patiently answered every question in every email I sent him. Through his explanations I was able to understand Generalizability Theory thoroughly. He patiently answered all of my questions, even the most naive ones. I wish him and his family all the best during his final journey. His contribution to my study was invaluable. I would like to thank Dr. Jerome Sheahan, from School of Mathematics, Statistics and Applied Mathematics, who patiently helped me to develop my background in statistics.

I certainly couldn't have done this without David Cunningham, Director/Chief Operations Officer at QPERCOM for the opportunity to work with the OSCE Management Information System, who gave access to the system. Without access to this software, it would not have been possible to conduct this Communication Skills in OSCE research. I thank my fellows at QPERCOM, Domhnall Walsh and Enda Griffin, for all the fun we have had during my PhD journey.

To my brothers and sisters, Ayu, Randy, bu Utami, and my friends and families whom I cannot mentioned all, thank you for your contributions in your own way toward my accomplishment.

Special thanks to my parents for their relentless praying for my success. To my mother-in-law, my greatest thanks to you for supporting, loving my family, and caring for the kids during my absence from Jogja for more than three years. My prayers for my late father-in-law, who gave me inspiring lessons “there is no such thing as stupidity or cleverness. Whether one is willing to study or not is what all that matters “

Finally, I would like to thank my wonderful wife, Sentagi Utami, who single-handedly took care of our kids throughout my study. You were always there for me when I needed you and were willing to listen to my stories. Distance and a six hour time difference were no match for our love. And last but not least, for my kids – Atiya, Daffa, Akila, and Danesh – you guys are the only reason I survived these 3.5 years.

Abstract

Communication skills (CS) are important skills that have a major impact in healthcare outcomes. These outcomes include, for instance, increased patient understanding, patient satisfaction and better pain control. With forty years of professional experience, a general practitioner will usually have conducted between 120 to 160 thousand clinical consultations. In fact, most of the decisions made to sue physicians arise from communication problems rather than from the original illness.

As in medical education settings, CS training invariably has a positive impact on student performance. An Objectives Structured Clinical Examination (OSCE) is a common measurement tool used to assess student CS. Concerning the OSCE, it has been reported that there has yet been no agreement reached amongst medical educators on the standard instrument to be used. The OSCE instrument serves as the station checklist. In order to draw meaningful interpretations of the assessment results, an instrument needs to be valid and reliable. Additionally, assessing CS is more difficult than assessing clinical skills. Furthermore, the lack of any standards in measuring CS precludes comparison of outcomes across assessment settings. In order to be able to compare, a standardised measurement is required.

This thesis has four distinct parts, in which further explanations are given in different chapters in order to address specific issues. The chapters comprise published or submitted papers. The first part is a systematic review, which aims to identify a reliable measurement instrument for assessing CS using the OSCE in undergraduate medical education. For this study, four databases were systematically searched (Pubmed, Embase, PsycINFO, and the ProQuest Education Databases) up until 2013. Studies that did not report both reliability and validity were excluded. This peer reviewed literature study identified those papers focusing on generic communication skills, history taking, physician-patient communication, interviewing, negotiating treatment, information giving, empathy, as well as 18 additional domains (ICC -0.12 – 1). There was only one real measurement instrument that was reported in more than one study, the MAAS-Global. The result of this systematic review supports previous research, the majority of the instruments were developed locally, and were only used or reported in a single study. There is a general lack of agreement amongst medical educators and researchers concerning the instrument to measure CS.

The second part of the study addresses the problematic situation in which many different measurement instruments exist, thus creating a dilemma when comparing student CS performance. A calibration process is being undertaken in order to achieve comparable results. The instruments, which are the OSCE

station checklists, were calibrated using an existing standardised instrument. 280 checklists from four disciplines contributing to Year Four OSCEs were analysed by three raters. The term calibration in this study, is the method for rating how close the items in the station checklists corresponded to the MAAS-Global. The reliability of the calibration process was analysed according to the Generalizability Theory. The result shows that the G-Kappa was 0.80. For two raters G-Kappa was 0.72. Nearly half the checklist items corresponded to MAAS-Global section 3, whilst 12 percent corresponded to section 2, and 8.2 percent to section 1. Thirty four percent of the items were not considered to be CS. This result confirms that the calibration of the OSCE station checklists is reliable. Further results also reveal that using two raters for the calibration remains reliable and is sufficient.

After the study proved that calibration is possible, the third part of the thesis sought for characteristics of the assessment from different disciplines. In this study a total of six circuits of three academic terms from Year Four OSCEs from two disciplines (General Practice and Psychiatry) were analysed. The proportion of section 3 items of the MAAS-Global (medical content) was larger than the proportion section 1 (sequential) and 2 (generic) items. General Practice stations scores (scale 1-100) ranged from 58 (sd=6) to 64 (sd= 6). MG scores ranged from 44 (sd=4) to 48 (sd=5). The Psychiatric stations scores ranged from 61 (sd=8) to 70 (sd=10), whereas MG scores ranged from 59 (sd=8) to 66 (sd=8). The results show that the MG score in the Psychiatry stations was significantly higher than in the General Practice station. This higher score was a direct result of the higher MG proportion in the Psychiatry station. The MG score can be categorized as a standardised CS score. From this study it can be shown that a comparison of outcome measurement from different measurement instruments is possible.

Having a comparable CS score, the fourth part of the thesis examined the CS type – section 1, section 2 and section 3 of the MAAS-Global – that affects student CS performance. As the number of international students increases over time, the students were divided into two groups: first, those who speak English as a first language only (L1); secondly, students who speak English as a second language (L2). According to the Working Memory (WM) theory, there are differences of function between L1 and L2. The student WM plays an important role in cognitive processes, by combining storage and analysis of information. WM is required in cognitive processing to undergo storage and analysis function. The WM of L2 students was apparently occupied in decoding unfamiliar language rather than solving the main task, hence reducing cognitive ability.

A retrospective study analysed Year 1 to Year 4 OSCEs from three academic terms and from five different disciplines. The reliability of the OSCEs ranged from $G = 0.28$ to $G = 0.79$ with a median of $G = 0.62$. The findings show that the majority

of CS assessments in OSCEs incorporate 50 percent or more of CS that focused on the MAAS-Global section 3 (medical content of the consultation). Of the OSCEs considered as section 3, nearly half of the CS assessments were considered as section 1 (CS occurred in the logical order of consultation) and half of those were section 2 (the general CS that occurred in all phases of the consultation). The results show that only in Year Four OSCEs, the L1 students performed significantly higher compared to the L2 students or General Practice stations by ($F(1, 368) = 21.46, p < .001$), and also for Psychiatry and Paediatrics stations ($F(1, 368) = 72.94, p < .001$) and ($F(1, 363) = 8.72, p < .003$), respectively. No significant differences were found for Obstetrics & Gynaecology stations.

It seems possible that these results are due to the CS assessment that focused on the medical content, which was then combined with the complexity of the case to be solved by the students in the Year Four OSCEs, creating disadvantages to the L2 students. The results of this study show that MG characteristics might have influenced student performance in CS assessment during OSCEs. Furthermore, an unanticipated finding was that for five stations, OSCEs never attained reliable assessment ($G=0.80$). In order to be able to produce reliable OSCEs, increasing the number of stations or increasing the number of examiners is recommended.

The thesis' discussion links all the papers with the main message: that a good assessment should use a standardised instrument in order to achieve a meaningful result. Some practical suggestions to increase the reliability of CS assessment in OSCEs are also presented in the thesis discussion.

Chapter 1. General Introduction

1.1. Introduction

In measuring, a proper measurement instrument should be used in order to obtain valid and reliable results. However, since first mentioned by Boon and Stewart in the late 1980s, there is still no general agreement amongst researchers and medical educators concerning an instrument to assess communication skills (CS) [1]. In 1999, the Association for Medical Education in Europe initiated a change from opinion-based education towards evidence-based education [2]. Evidence-based education is the practice of using the latest best practices in medicine and education in the health sector and incorporating that theory into an education process, thus, it is called Best Evidence-based Medical Education (BEME) [2,3].

The research presented in this thesis aims to further improve the field of CS assessment in undergraduate medical school. Furthermore, it focuses on using a specific assessment tool, namely Objective Structured Clinical Examination (OSCE).

This thesis presents several studies regarding the assessment of CS in OSCEs in undergraduate medical education. This introductory chapter explains the context and rationale of the thesis. It also includes some background information relating to CS and assessment in undergraduate medical education.

At the end of Chapter 1, a synopsis of the general aim of the thesis and the research questions derived from the aim is provided.

1.2. Rationale

1.2.1. Communication Skills

CS is an important component in medical practice. Many studies indicate that good communication will improve patient satisfaction and patient outcomes.

During the typical 40 years of a general practitioner's career, he or she will conduct between 120,000 and 160,000 clinical consultations, and it is clear that CS plays an important role in the consultation process [4]. It is also worth noticing that a patient's decision to sue a physician is most likely not due to the original illness, but to the poor handling of communication between doctor and patient. One study even suggests that 80% of the cases involved are due to communication issues, rather than the quality of care [5]. In the past decade, patient safety has become the most critical component of the healthcare system. Therefore, to eliminate the problems caused by a lack of information received by the patient, good practice and effective CS will ensure that the patients adequately understand their diagnoses and the management of their care plans.

Many forms of communication occur in the clinical consultation environment (i.e. verbal, non-verbal and written). The consultation process itself can be divided into several phases that occur sequentially: the introduction, a request for help, a physical examination, diagnosis, management and the evaluation of the consultation. It is understandable that verbal and non-verbal communication have important roles in clinical consultation. Building and maintaining a relationship with the patient will ensure that history-taking is facilitated in a smooth, efficient manner, thus explaining the diagnosis with empathy, so that the patient will understand the management of the treatment. This, in turn, will increase the patient's commitment.

CS have a positive impact on students' performance during the education process [6]. For the students, building CS also increases awareness of time management, and will have a direct impact on the effort to gain better skills in assessing the patient.

1.2.2. OSCE

The Objective Structured Clinical Examination (OSCE) is an assessment tool. This tool was developed by Harden, Dundee University, and introduced in late 1970

[7]. OSCE consists of several stations, and each station assesses different skill sets. The student will progress from one station to the next in sequential order. At each station, students are given a certain time limit (e.g. 5, 10 or 15 minutes) to undergo an assessment that is specifically designed for that station. If the OSCE setting, for example, has 10 stations, the student in that examination needs to pass through all of those 10 stations OSCE is commonly used to assess the student's ability in regard to clinical skills in the medical education environment.

Miller introduced the model known as Miller's pyramid in 1990, which illustrates which assessment tools are suitable for learning outcomes [8]. In this model, as shown in Figure 1, students commence their learning process by building knowledge, and will continue towards the top level of learning, which focuses on learner behaviour. A written assessment is a suitable tool to explore students' 'know' and 'know how' levels. Additionally, in the level 'show how', clinical and practical assessment tools are needed to explore their knowledge and how the students execute it in a controlled and practical environment. OSCE has been, up to now, the most suitable tool for this particular assessment. Meanwhile, appropriate tools to assess the practical 'does' level are observation, portfolios, logs and peer assessment.

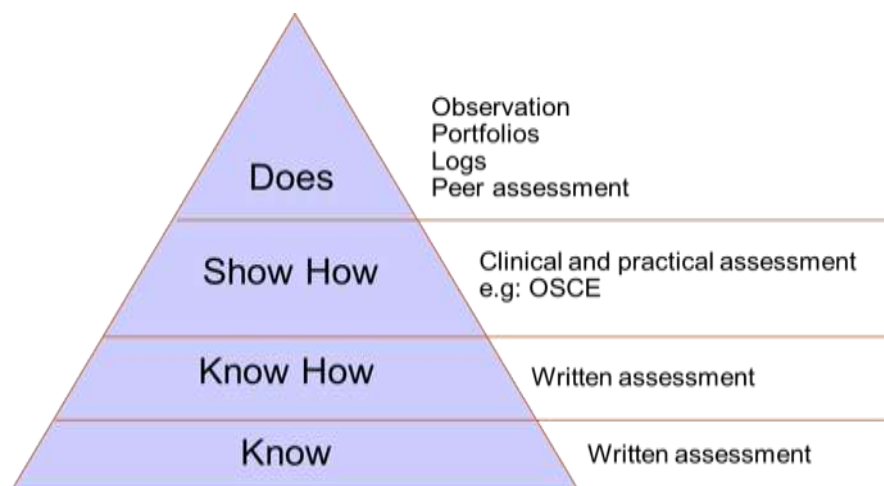


Figure 1.1: Miller's Pyramid.

1.2.3. The CS Assessment

First and foremost, it is important to know how the student's CS increases during the course of the education process. Students need to reflect on their own ability in regard to CS. Meanwhile, medical educators need to know whether the learning outcomes meet the learning objectives, and where further improvement is needed, based on the results of the examination, in addition to comparing different cohorts of students to understand the phenomenon in academic terms.

CS cannot be separated from clinical skills, since communication skills in the clinical environment are bound by context and content. The OSCE provides the closest conditions to the real conditions faced in the professional career of a health care worker. This means that the assessment of CS must be aligned with the study objectives, learning outcomes and, most importantly, the assessment of the clinical context [9]. In assessments, validity and reliability are mandatory. Validity means that the assessment is measuring what it is supposed to measure, while reliability means that the instrument is able to retrieve similar results repeatedly. One author has observed that there cannot be reliable measures if they are not valid, and vice versa; therefore, an assessment must be both valid and reliable [10].

There are two types of assessments: summative and formative. The easiest way in which to distinguish between the two is by considering the whole process of education [11]. Formative assessment occurs during the process of learning, and provides the students with feedback from which to learn their strengths and weaknesses. Summative assessment, meanwhile, is made at the end of the learning process, which allows the educational institution to judge whether the student already meets the requirements for a certain profession – such as semester or year-end assessments up the licensing exam by the professional bodies. Formative assessment occurs repeatedly during the education process.

As regards a general ‘rule of thumb’ in assessing the CS of students, to date there is no agreement in relation to a widely used assessment instrument [9]. The most important aspect to consider is the need to divide the assessment into various levels of competence: basic and applied knowledge (knowledge), performance in simulation (skills), and action in the workplace (attitude). Basic and applied knowledge can be appropriately assessed with a written test, while performance in a near-real professional condition is best assessed with an OSCE. Finally, performance in the workplace is commonly measured using so-called ‘workplace-based assessments’.

1.2.4. CS Measurement Instrument

CS in clinical skills is commonly assessed with an OSCE. In terms of measuring, the one important requirement is that the measurement instrument has to be standardised. This means that the instrument has to be valid and reliable. There are several instruments that were designed to measure CS (e.g. RIAS, SEGUE, MAAS-Global) [1,12,13]. These instruments can be used for real-time assessment or recorded/video-taped sessions. SEGUE, for example, is an instrument that was designed to measure a recorded session. MAAS-Global is an instrument that was designed for direct observation, while other instruments to assess CS have been used by patients or standardised or simulated patient [12]. Since the late 1990s, researchers have been warning that there is no agreement amongst researchers and medical educators concerning a standard measurement instrument to assess CS [1,14]. The common practice is for educators to use existing instruments that have been developed to assess a physician’s CS, modified in some way to fit the purpose of assessing students. The MAAS-Global, for example, is a measurement instrument used to assess a physician’s CS, which has been used by several universities to assess the CS of medical students [15]. There is a reminder by Silverman (2007) that, in order to assess CS, a proper measurement instrument is needed in order to achieve a reliable result [16].

It is interesting that some studies have reported a low OSCE reliability coefficient [17]. The low reliability of an assessment means that there are many errors in the results. According to the Classical Test Theory (CTT), a score that is generated from assessments is considered an Observational Score, which is actually a summary of the True Score and Error. Whenever an examination has low reliability, this means that a great deal of variance (e.g. error, noise) was generated around the 'true score', and it is more difficult to know the true ability of the students. The observed score is just one score out of a universe of scores [18]. In OSCEs, a specific reliability analysis method is suggested as standard by the Association of Medical Education in Europe (AMEE), the Generalizability Theory (G-Theory) [19]. This method is used to ensure that the correct calculation is conducted to analyse the reliability in a way that is suitable to the nature of OSCE. Furthermore, G-Theory not only calculates reliability, as with the CTT method, but is also able to pinpoint the source of error, and to quantify it [18]. Another useful feature of G-Theory is the ability to predict the hypothetical design of the examination model, which enables the improvement of future examinations.

Language tests closely resemble CS measurement; both assess communication ability (verbal and written). However, language tests are already well-developed in terms of measurement instruments. There are several standardised instruments – i.e. TOEFL, IELTS, TOEIC – that enable comparisons amongst different groups, regardless of location. For example, a score of 500 in a TOEFL paper-based test in one country is always better than a 370 score for the same test in another country. These results are standardised and comparable. In contrast, researchers and medical educators are still unable to agree on a standard instrument for CS assessment [14]. A simple analogy for this condition is the attempt of measuring the height or length of a chair and table using a rubber band. One could claim that both have the same length, which is twice the length of the rubber band, but care needs to be taken in terms of how we hold

and stretch the rubber band. Measuring length with a rubber band will render an unreliable result.

1.2.5. Calibration of Measurement Instrument

As already mentioned above, an unstandardised measurement instrument creates challenging situations in terms of comparability between cohorts of students, individual students, different settings of OSCEs, or wider groups of students, such as comparison between institutions. We must consider how complicated the procedure is when comparing two cohorts of students from different medical schools. In order to be able to compare the CS of students from different schools, the measurement instruments have to be standardised. The assessment results have to be able to be compared on an equal footing; hence, both institutions have to use the same measurement instrument. In this case, the MAAS-Global was chosen. Since only one institution was familiar with assessment using the MAAS-Global, the other institution had to undergo training for the examiners to be able to use the instrument correctly. Moreover, the students who had already been assessed with this instrument had advantages over the students who were not assessed with this instrument. The point of this explanation is simply to compare the CS competence of students from different medical schools, and there was a great deal of effort involved, which was also time-consuming and cost-intensive. Had there been a standardised measurement instrument in place, the task could have been completed more conveniently, and in a simpler manner.

1.2.6. The Effect of L1 and L2 in Students Performance

The term 'L1' refers to the condition whereby students speak and communicate in an educational institution where the language of instruction is their first language. 'L2' refers to students who have to communicate using a language other than their first language. In the past two decades, there has been evidence that many students are choosing to study and enrol in higher education involving

languages different to their first language. An obvious concern, in this regard, is the potential difficulties and challenges associated with language proficiency and academic performance [20]. There are no additional difficulties associated with the L1 language; however, L2 students can face problems regarding their study performance. The reason for this is that there is more to language proficiency than mastering reading and writing, especially in a clinical environment, where students have to interpret language in the context of non-verbal cues and cultural values [21]. In OSCEs, with standardised patients (SPs) communicating in the local language, L2 have difficulties in understanding societal norms, and sometimes also experience discomfort when asking about intimate matters [22]. The result of these problems is difficulties in communicating with patients, and the result of this is poor performance in clinical settings.

According to the Working Memory (WM) Model theory, any studies that deal with differences between L1 and L2 should consider this model. WM works by receiving information, storing it in Short Term (ST) memory, accessing Long Term (LT) memory and, most importantly, making an analysis, reasoning and reaching a decision. This analytic and phonetic ability is unique to the WM, and therefore it has a greater role in cognitive activities [23]. WM has three different aspects of information gathering, which are verbal information, spatial information and non-verbal information. When the L2 receives information, WM tends to save the unfamiliar language in the ST memory and seek to access the LT memory. Because of the fact that, in communication, cues or signs are more difficult to understand, and terms are easier to memorise, the analytical process of cues will take longer for an L2. Even then, the production of verbal output by an L2 takes longer than that by an L1 [24]. While there is much hesitation amongst medical educators about judging students' performance in relation to language ability, since it could be interpreted as racist, with the understanding of the process of how the brain works (WM theory), this hesitation should not be an issue at all. It is clear that it is not particular to language or culture; rather, the brain takes

longer to process information. The educational institution should consider openly recognising the problem that international students face in the process of learning, and provide a solution that bridges the gap between local and international students. It is suggested that the school could provide training in the use of the language for medical purposes and provide extra activities in order to improve the foreign students' social skills.

1.3. General Aim, Research Questions, and Outline of the Thesis

The general aim of this research study is, therefore, to investigate the availability of a standard measurement instrument to assess the CS of undergraduate medical students in OSCEs and, furthermore, to explore the use of a standardised CS measurement instrument. Five research questions have been developed in order to provide answers to support the general aim of the study. The decision on a 'gold standard' measurement instrument should be based on two main criteria: validity and reliability. More evidence needs to be explored in order for a CS measurement instrument to be accepted by the medical education community. Although there are several good options, we have to remember that acceptance of the new concept will not be easy. Furthermore, the acceptance of the result of the study in the 'real' environment will take time. Situations are required whereby educators and researchers can collaborate in the development and implementation of the proposed method. In this study, the author actively collaborated with other researchers from different disciplines, and even different universities.

To begin the study, the author needed to search for evidence of existing instruments which were already being developed and used to assess the learning process of undergraduate medical students. This was formulated via the first research question:

- 1. Is there any agreement amongst researchers and medical educators on a measurement instrument to assess the CS of undergraduate*

medical students in OSCEs? (Chapter 2)

It has been suggested since the late 1990s that researchers and medical educators need to work on a standard measurement instrument [1,14]. However, to date, there has been little agreement on a standard CS assessment instrument. Moreover, Brannick (2009) considers assessing CS to be difficult [17].

Since little is known about a specific CS measurement instrument that can be used in OSCEs to assess medical students, a systematic review was conducted. This study attempts to find a common measurement instrument that has been used amongst researchers and medical educators around the world. It seeks to identify any 'gold standard', determining whether there are any standardised instruments akin to the language tests. With this review, it is expected that solid evidence of what instrument should be used, and what aspects need to be considered when we assess CS, will emerge. If the study identifies a valid and reliable measurement instrument, there must be some scope for comparison as to whether the assessment of CS in our institution's OSCEs is different from the findings of the systematic review. It remains to be seen whether we can convert the instrument's scale, as with converting from Celsius to Fahrenheit. An experiment involving the calibration of our own OSCE station checklists is needed. The possibilities of building a new method to calibrate the rubrics have led to the second research question:

2. *What is the method used to calibrate the existing CS assessment of the OSCEs, which have different standards? (Chapter 3)*

A study was conducted in order to explore the extent to which we can develop a method of calibrating OSCE checklists. In this process, a careful examination of all Year Four OSCE checklists from three academic terms (2010/2011, 2011/2012 and 2012/2013) was conducted. The Year Four OSCEs were administered by four disciplines: Obstetrics & Gynaecology, Paediatrics, Psychiatry and General

Practice. Based on previous research detailed in Chapter 2, it was found that the MAAS-Global was the most suitable candidate with which to commence. There might be debate regarding the standard to choose, but the most important factor in measuring is using a standardised instrument. An analogy here involves measuring the height of a table or chair, deciding whether to use centimetres or inches, depending on the scale.

The calibration method in this study needed to be highly reliable. Another aspect to consider was how complicated and time-consuming it would be. It was known that a faculty member would not have the luxury of extensive time for the project, and so this method was also required to meet the criteria of being simple and requiring as few working hours as possible. Nonetheless, it had to produce a reliable result. At this stage, the study should have been able to identify a type of 'gold standard' in regard to a CS measurement instrument, and then to apply the standard into existing standards in different settings of OSCEs. The next step involved assessing the results of the OSCE of the students. Therefore, the third research question was as follows:

3. *Are there any differences in CS assessed from different settings of OSCEs? (Chapter 4)*

Based on a standard used in a previous study (as detailed in Chapter 3), the calibration result involved the mapping of different types of CS items. The types of CS items are based on the MAAS-Global, and the way they determine various OSCE station forms is called the MAAS-Global characteristic (MG characteristic) of that specific station. In student assessment, this approach opens a new level of understanding of the existing assessment. In addition, many studies did not clearly report what standard they used [13,14].

In Chapter 4, this study explores the CS based on the MAAS-Global as the standard, whilst also addressing its effect on the overall OSCE score.

Furthermore, various medical disciplines contribute to OSCE with discipline-specific OSCE stations. This study reveals the type of CS used by each participating discipline in their specific stations and explores the effects on the overall CS score. At the end of this chapter, a comparison of CS scores achieved in two integrated modules and different settings of OSCEs is possible. It is as easy as comparing a TOEFL or IELTS score. Since this study is focusing on CS, and considering the fact that many universities have international students, it is expected that there are differences between local and international students' performance in CS. Therefore, the fourth research question that arose was the following:

4. *Is there any relationship between the type of CS affecting the performance of students when studying in a foreign language environment (L2)?* (Chapter 5)

It has been demonstrated in previous studies that language proficiency correlates with the students' performance. While CS in medical schools is highly associated with the content and context of clinical problems, it is highly likely that CS will have different effects on local and international students. The local students will face less of a challenge, since they speak their first language (L1) throughout their learning process. Based on the Working Memory model, the analytical processes of the L1 and L2 are different. Within this study, it is expected to find a correlation between each domain of CS and how the brain processes simultaneous information, which will have an effect on the students' performance, particularly in OSCEs.

Along with the analysis of many OSCEs in Year One (1MB), Year Two (2MB), Year Three (3MB) and Year Four (4MB), this paper will determine how effective the examination is in terms of reliability. Therefore, the fifth research question was formulated as follows:

5. *How many OSCE stations will produce a reliable examination?* Chapter 5)

At the conclusion of any assessment, the reliability of the OSCEs is one important variable that must be considered. A lower reliability number signifies more error in the assessment, and therefore it is difficult to know the 'true' ability of the students. As Cook et al. (2006) have mentioned, reliability should be reported in any OSCEs study, although the present study found that some studies use reliability analysis that is not recommended by medical education associations, or else they did not report reliability [10,19,25].

The fifth chapter explores and analyses all OSCEs that have been using the OSCE Management Information System (OMIS) across three academic terms (2010/2011, 2011/2012 and 2012/2013). The different settings of OSCEs from five disciplines which were using OMIS will, hopefully, reveal the best setting to be considered for future OSCEs. Finally, Chapter 6 summarises and interprets the main findings of this thesis in regard to the context, answering the five research questions presented above. The implications of these research findings, as well as suggestions for future research in CS assessment and OSCE in medical schools, will also be detailed. It should be noted, in any case, that most of the chapters of this thesis have been published as articles or submitted to journals, and therefore, some duplication of material in each chapter is inevitable.

1.4. References

- [1] Boon H, Stewart M. Patient-physician communication assessment instruments:: 1986 to 1996 in review. *Patient Educ Couns* 1998;35:161–76. doi:10.1016/S0738-3991(98)00063-9.
- [2] M. Harden R. Best evidence medical education: the simple truth. *Med Teach* 2000;22:117–9.
- [3] R. Hart I. Best evidence medical education (BEME): a plan for action. *Med Teach* 2000;22:131–5.
- [4] Lipkin M, Putnam SM, Lazare A. *The medical interview*. Springer; 1995.
- [5] Levinson W. *Physician-patient communication: A key to malpractice*

- prevention. *JAMA* 1994;272:1619–20. doi:10.1001/jama.1994.03520200075039.
- [6] Yedidia MJ, Gillespie CC, Kachur E, Schwartz MD, Ockene J, Chepaitis AE, et al. Effect of communications training on medical student performance. *JAMA J Am Med Assoc* 2003;290:1157–65. doi:10.1001/jama.290.9.1157.
- [7] Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J* 1975;1:447–51.
- [8] Shumway JM, Harden RM. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Med Teach* 2003;25:569–84. doi:10.1080/0142159032000151907.
- [9] Kiessling C, Essers G, Anvik T, Jankowska K. General principles for the assessment of communication skills. (http://www.each.eu/wp-content/uploads/2014/07/General_principles_for_the_assessment_of_communication_skills_final.pdf):(accessed 4.26.2015).
- [10] Cook DA, Beckman TJ. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *Am J Med* 2006;119:166.e7–166.e16. doi:10.1016/j.amjmed.2005.10.036.
- [11] Fred C. Removing confusion about formative and summative evaluation: Purpose versus time. *Eval Program Plann* 1994;17:9–12. doi:10.1016/0149-7189(94)90017-5.
- [12] Schirmer JM, Mauksch L, Lang F, Marvel MK, Zoppi K, Epstein RM, et al. Assessing communication competence: a review of current tools. *Fam Med* 2005;37:184–92.
- [13] Ong LML, De Haes JCJM, Hoos AM, Lammes FB. Doctor-patient communication: a review of the literature. *Soc Sci Med* 1995;40:903–18.
- [14] Beck RS, Daughtridge R, Sloane PD. Physician-patient communication in the primary care office: a systematic review. *J Am Board Fam Pract* 2002;15:25–38.
- [15] van Es JM, Schrijver CJW, Oberink RHH, Visser MRM. Two-dimensional structure of the MAAS-Global rating list for consultation skills of doctors. *Med Teach* 2012. doi:10.3109/0142159X.2012.709652.
- [16] Silverman J. The Calgary-Cambridge guides: the “teenage years.” *Clin Teach* 2007;4:87–93.
- [17] Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ* 2011;45:1181–9. doi:10.1111/j.1365-2923.2011.04075.x.
- [18] Cardinet J, Johnson S, Pini G. *Applying Generalizability Theory Using Edug*. Taylor & Francis; 2012.
- [19] Bloch R, Norman G. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Med Teach* 2012;34:960–92. doi:10.3109/0142159X.2012.703791.
- [20] Harvey T, Robinson C, Frohman R. Preparing culturally and linguistically diverse nursing students for clinical practice in the health care setting. *J Nurs*

Educ 2013;52:365–70. doi:10.3928/01484834-20130529-02.

[21] Crawford T, Candlin S. A literature review of the language needs of nursing students who have English as a second/other language and the effectiveness of English language support programmes. *Nurse Educ Pract* 2013;13:181–5. doi:10.1016/j.nepr.2012.09.008.

[22] Couper J, Hawthorne L, Hawthorne G, Tan E-S, Roberts A. Communication skills and undergraduate psychiatry: a description of an innovative approach to prepare Australian medical students for their clinical psychiatry attachment. *Acad Psychiatry* 2005;29:297–300.

[23] Sawyer S, Eschenfelder KR. Social informatics: Perspectives, examples, and trends. *Annu Rev Inf Sci Technol* 2002;36:427–65. doi:10.1002/aris.1440360111.

[24] Kormos J, SáFáR A. Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Biling Lang Cogn* 2008;11. doi:10.1017/S1366728908003416.

[25] Casey PM, Goepfert AR, Espey EL, Hammoud MM, Kaczmarczyk JM, Katz NT, et al. To the point: reviews in medical education--the Objective Structured Clinical Examination. *Am J Obstet Gynecol* 2009;200:25–34. doi:10.1016/j.ajog.2008.09.878.

Chapter 2. Reliability and Validity of OSCE Checklists Used to Assess the Communication Skills of Undergraduate Medical Students: a Systematic Review

This chapter was published as a paper in the Patient Education and Counseling, 2015, <http://dx.doi.org/10.1016/j.pec.2015.06.004>

2.1. Abstract

Objectives. To explore inter-rater agreement between reviewers comparing reliability and validity of checklist forms that claim to assess the communication skills of undergraduate medical students in Objective Structured Clinical Examinations (OSCEs).

Methods. Papers explaining rubrics of OSCE checklist forms were identified from Pubmed, Embase, PsycINFO, and the ProQuest Education Databases up to 2013. Included were those studies that report empirical validity or reliability values for the communication skills assessment checklists used. Excluded were those papers that did not report reliability or validity.

Results. Papers focusing on generic communication skills, history taking, physician-patient communication, interviewing, negotiating treatment, information giving, empathy and 18 other domains (ICC -0.12 – 1) were identified. Regarding the validity and reliability of the communication skills checklists, agreement between reviewers was 0.45.

Conclusions. Heterogeneity in the rubrics used in the assessment of communication skills and a lack of agreement between reviewers makes comparison of student competences within and across institutions difficult.

Practice implications: Consideration should be afforded to the adoption of a standardized measurement instrument to assess communication skills in undergraduate medical education. Future research will focus upon evaluating the potential impact of adoption of a standardized measurement instrument.

2.2. Background

Physicians' communication skills (CS) have a considerable impact upon quality of health care, whereby good CS improve healthcare outcomes, such as physiologic status, pain control, and emotional health, and significantly increase patient understanding and patient satisfaction [1, 2].

Effective physician-patient communication is essential in ensuring that patients adequately understand their diagnoses, treatment options, medications, plans for referral and prognosis. Dissatisfaction with physician-patient communication is known to be a leading factor influencing patients' decisions to initiate medical negligence proceedings [3,4]. Existing research demonstrates that errors in physician-patient communication include inadequate information-giving, reluctance to adopt a specific partnership style, being in a hurry and failing to respond to patients' feelings [5-7]. In clinical settings, CS take verbal, nonverbal and written forms. From the point of view of physician-patient interaction, CS can be classified according to purpose. For instance, initiation of a session, gathering information, providing structure to the interview, building a relationship, explanation and planning, closing the session and other specific issues [2, 4, 8, 9].

Within medical educational settings, practical CS training has been shown to improve medical student performance in relationship building, time management and patient assessment [10]. According to Humphris, 2002, medical students' acquisition of CS is influenced not only by structured teaching sessions but also by incidental learning. The development of communication knowledge has a small, but significant, influence on performance [11].

The Objective Structured Clinical Examination (OSCE), an assessment method introduced by Harden in 1975, is the assessment tool most commonly used for assessment of clinical skills in undergraduate medical education [12]. Research

suggests that the OSCE is appropriate for high-stakes assessment [13]. In addition to practical clinical skills, the OSCE can be used to assess complex CS [14]. Such assessment may take the form of OSCE stations dedicated to the assessment of CS or stations testing specific subject areas or domains of CS alongside other clinical skills. In our Medical School, the majority of OSCE stations combine both the assessment of domains of CS with assessment of a specific set of clinical skills. It is acknowledged that the combination of CS domain checklists with clinical skills checklists is likely to influence the choice and design of the assessment tools. Interpretation of student performance in such stations can be complicated by the combination of CS and clinical skills assessment, such that students may compensate between these skills to achieve a pass grade overall, whereas their performance in the individual competencies is often not immediately apparent. The OSCE itself has evolved into many variations, the Objective Structured Clinical Assessment (OSCA) and the Group Objective Structured Practical Examination (GOSPE) [15-18].

Very many different measurement instruments have been used to evaluate CS in OSCEs [19]. To examine such skills, two types of scale ratings are frequently used. The first type is that of a “behavioral checklist” and the second is a “multi-point global scale/global rating scale” (GRS). There is evidence supporting the use of global rating scales (GRS’s) rather than checklists [15]. Research suggests that GRS’s have higher internal consistency when compared against checklists, and furthermore, that using both GRS’s and checklists in combination can improve content validity [20, 21].

CS can either be assessed during real time assessments or after a recorded session. Those collecting data in real time have the potential advantage of being able to provide instant feedback to participants, whilst recorded sessions have the advantage of generating permanent data that can be used for repeated analysis [15]. Regardless of the method used, it is important for medical

educators to evaluate students' CS on a number of occasions over their entire course of study so that an improvement in ability can be recognized and so that those students who are failing to progress can be identified [22]. However, heterogeneity in measurement instruments used to assess CS in OSCEs limits the comparability of student performance between examinations settings. It would be expected that most institutions would be using similar rubrics for assessment of CS in different years of their degree programmes, thus allowing easy comparison between students and of individual progress across academic years. Ideally, there would also be consistency within and between institutions in terms of the rubrics used to assess CS.

There is an existing body of research pertaining to the assessment of CS. Beck et al (2002) reviewed measurable verbal and non-verbal CS of physician-patient communication, Ong et al (1995) compared interaction analysis systems and Boon and Steward (1998) reviewed available instruments to assess physician-patient communication [2, 4, 19]. Schirmer et al (2005) compared to what extend the instruments measured essential elements of communication in the family medicine context [23]. The aim of the present review is to explore inter-rater agreement between reviewers analyzing quality and content of papers systematically by comparing whether reliability and validity of checklist forms that claim to assess the CS of undergraduate medical students in OSCEs are described appropriately in these papers. Agreement between raters about quality and content of the included papers is expressed in an intra class correlation coefficient (ICC).

2.3. Method

A preliminary narrative literature review, pertaining to clinical CS and OSCEs, was conducted by the Principle Investigator (PI), WS, in order to ensure that key points and conceptual frameworks were adequately covered in later search strategies. A list of keywords was developed from the results of this exercise, so

that they could form the basis for a more extensive literature search detailed below.

A search was performed in order to identify studies which were published between January 1975 (first description of the OSCE) and December 2012, in peer reviewed publicly available international journals published in English. The following databases were searched: PUBMED, EMBASE, PsycINFO Ovid, and ProQuest Education Databases (consisting of ERIC, British Education Index, and the Australian Education Index).

Boolean operators (i.e. AND, OR, NOT or AND NOT) were used as conjunctions to combine or exclude keywords in a search, thereby resulting in more focused and relevant results in PUBMED. These were adapted accordingly for the other databases. The examples of search terms identified in this manner were “Objective Structure Clinical Examination”, “OSCE”, or any variation of OSCE including the abbreviations. This was followed by combining results, using the Boolean logic AND, with words from communication domains such as “communication”, “history taking”, “physician-patient relationship”, “interview” or “counseling”.

A series of search strategies was utilized to ensure correct results and limits were applied to remove false results. The search strategy for PUBMED is provided below. This was adapted accordingly for the other three databases.

Table 2.1: Steps conducted in the initial narrative review to retrieve the appropriate literature for the systematic critical appraisal of the literature.

1.	OSCE
2.	Objective Structured Clinical Examination
3.	OR 1-2
4.	MMI
5.	“multiple mini interview”
6.	“multiple mini-interview”
7.	OR 4-6
8.	MiniCex

9.	Mini-cex
10.	"mini Clinical Evaluation Exercise "
11.	mCEX
12.	OR 8-11
13.	OSCA
14.	"objective structured clinical assessment"
15.	OR 13-14
16.	TOSCE
17.	"team observed structured clinical encounter "
18.	OR 16-17
19.	GOSPE
20.	"group objective structured practical examination"
21.	OR 19-20
22.	3 or 7 or 12 or 15 or 18 or 21
23.	Communication
24.	"communication skills "
25.	"history taking"
26.	Consultation
27.	"consultation skills "
28.	"breaking bad news "
29.	"cross cultural "
30.	"interpersonal relation "
31.	"end of life"
32.	"informed consent"
33.	Anamnesis
34.	Interview
35.	"medical interview"
36.	"doctor-patient interaction"
37.	"doctor-patient relation"
38.	"physician-patient relation"
39.	"physician-patient interaction"
40.	Referral
41.	Counseling
42.	"non verbal communication"
43.	"electronic communication"
44.	"email communication"
45.	"doctor-nurse communication"
46.	"physician-nurse communication"
47.	"health beliefs"
48.	"treatment plan*"
49.	OR 23-48
50.	22 AND 49

51.	Dentist*(Title/Abstract)
52.	Veterinary(Title/Abstract)
53.	Pharmacy(Title/Abstract)
54.	Pharmacist(Title/Abstract)
55.	OR 51 - 53
56.	50 NOT 55

Whilst the Boolean string operator “NOT” was applied in PUBMED, application to the other databases was problematic due to different Boolean logistics. Thus, we used reference management software, known as Zotero, to overcome this issue. The PI, WS, carried out a manual search of the references of identified studies in order to identify further relevant studies.

We included studies which described the assessment of CS using OSCEs in undergraduate medical students. Only papers referring to undergraduate medical students were included. Studies conducted within dentistry, veterinary, pharmacy and other para-medical disciplines were excluded. Papers were included if they described OSCE stations which were entirely dedicated to the assessment of CS. Papers which described OSCE stations that assessed CS only as a component of a broader assessment, such as clinical examination or procedural skills, were also included. Studies were excluded if they did not provide empirical validity or reliability information in relation to the assessment checklist used (i.e. papers had to explicitly state the validity and/or reliability of their assessment checklist or a reference to an existing study of the validity and/or reliability of the checklists). Studies were included regardless of the nature of the specific clerkship that the OSCE was associated with and regardless of whether or not the participating students originated from the same year of study.

For the second ‘systematic review’, we included all identified CS measurement instruments (checklists) used in OSCEs. Instruments that measured CS in

assessment types other than OSCEs were not included. Studies that did not provide a description of the CS measurement instrument were excluded.

Each reviewer analyzed the included literature using a data-extraction template¹. The template was designed using keywords and assessment rubrics found in potentially relevant papers. It consisted of 2 categories, whereby category one sampled 22 domains of CS as assessed by an examiner and category two sampled 5 domains of CS as assessed by a Standardized Patient rater (SP rater). Other information that was extracted from each paper included the study sample size (number of students), the duration of stations (recorded in minutes), the utilization or otherwise of a CS checklist and referral to any professional board or licensing bodies.

After an initial meeting to agree the meaning of each item on the data-extraction template, WS, TK, KK independently analyzed each research paper. Where two out of three reviewers were in agreement (initial agreement in percentage), these items were discussed with a view to achieving complete agreement where possible (resolved disagreement in percentage). To correct for chance an Intraclass Correlation Coefficient (ICC) was calculated. Data was entered into SPSS (version 20) and the levels of agreement between reviewers for each of the 27 domains of CS were measured using ICC. Full agreement between reviewers (ICC = 1) means 100% agreement on items assessed. No agreement (ICC = 0) means that reviewers did not agree at all on the items that were assessed. An agreement of 0.45 means reviewers agreed on 45% of the items that were assessed with a correction for agreement by chance.

Ethical approval was not required for this review.

¹ The data-extraction template was developed by collecting CS related key-words from 87 potentially relevant papers (see Figure 2.1). The result was 27 CS domains of data-extraction templates that consisted of 22 domains of CS which were assessed by examiners and 5 domains of CS which were assessed by standardised patient raters (SP rater).

2.4. Results

2.4.1. Search Results

The initial literature search identified 1,998 papers (Figure 2.1). After removal of duplicates, 1,358 papers remained. By review of the titles and abstracts, 613 were excluded on the basis of irrelevancy. A further 557 papers were excluded as they were not related to OSCEs in undergraduate medical schools. Manual review of the titles and abstracts of the remaining papers identified a further 20 duplicates and 13 non-English language papers, all of which were excluded. In cases where it was not possible to make the decision to include or exclude a paper based upon its title and abstract alone, the full text of the paper was reviewed. Review of the full text of the remaining 87 papers revealed that 48 did not assess CS and a further 5 did not provide validity or reliability data. All of these papers were excluded, thus leaving 34 papers to be included in the review (Table 2.1).

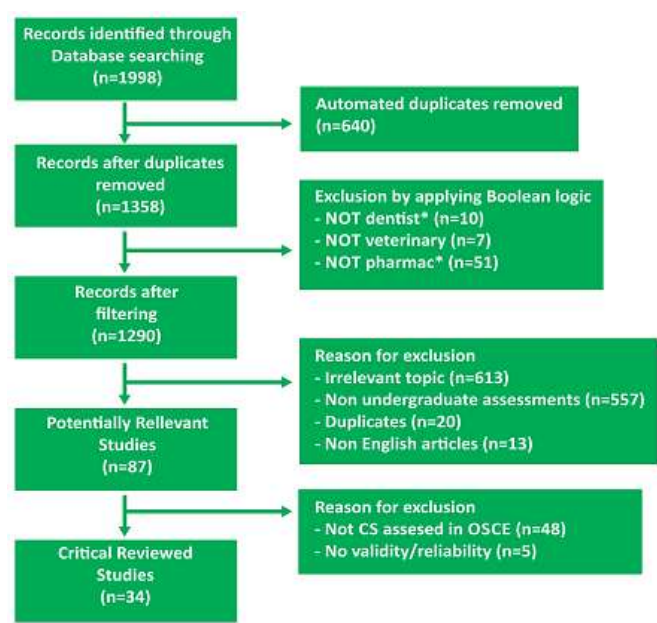


Figure 2.1: The different stage of the systematic review.

2.4.2. Content Analysis

The number of student participants in individual studies ranged from 36 to 476, with an average number of 185 students. Twenty five studies reported the duration of CS stations, with the shortest being 5 minutes long and the longest being 20 minutes long [11, 24-46], whilst almost half of the reviewed papers ranged from 5 to 8 minutes [11, 24, 26-31, 33, 36-39, 42-44]. Four studies reported short and long case scenarios with obviously different durations [11, 31, 32, 37], in contrast to nine studies which did not report station length at all [15, 47-55]. Sixteen studies reported references to validity and reliability studies of the assessment forms being used [11, 15, 24, 26, 27, 30, 39, 40, 42-46, 52-54], whilst six studies reported only validity [31, 33, 37, 50, 55], and twelve reported only reliability [25, 28, 32, 34-36, 38, 41, 47-49, 51]. Two studies compared assessments of medical student CS across different institutions [40, 52], whilst the remainder were based in a single institution. The included studies involved participants across the range of 1st year to final year. Sixteen studies report upon the assessment of 3rd year students [15, 25-27, 29, 31, 34, 36-38, 41, 47, 51-54], and six studies reported upon more than one year of study [11, 15, 31, 37, 40, 54]. In two studies students were assessed only by SP raters [50, 51]. In eight studies students were assessed by examiners and by SP raters [11, 26, 31, 34, 36, 45, 52, 53]. In the remaining 24 studies students were assessed by examiners alone.

The majority of the papers examined focused upon eight domains, which included generic CS, physician-patient communication, history-taking, focused history-taking, interviewing, negotiating plan/treatment, information giving and empathy (Table 2.2). The papers where SP raters were involved as assessors focused mainly on generic CS and interpersonal skills. The term “generic CS” was used where reviewed papers only mentioned 'communication skills' without

giving any additional information regarding specific descriptions of the CS domains being addressed.

Regarding adherence to recognized standards, three papers reported use of the Calgary-Cambridge Observation Guide (CCOG) [37, 48, 49], whilst two others used the Maas-Global and revised Maas-Global (Maas-R) [33, 40], and two papers used the Standardized Patient Satisfaction Questionnaire (SPSQ) [26, 31]. The Patient Perception Questionnaire (PPQ), the American Board of Internal Medicine Patient Satisfaction Questionnaire (ABIM PSQ), the Liverpool Communication Skills Assessment Scale (LCAS), the Global Simulated Patient Rating Scale (GSPRS) and the WHACS mnemonic which were each used only by a single study [11, 27, 31]. The WHACS mnemonic provides an essential checklist for history taking on occupational and environmental health and was created by the Environmental Medicine Curriculum committee of the South Carolina Statewide Family Practice Residency Program [27, 56]. Chessman (2003) and Humphris (2002) incorporated more than one recognized standard into their checklists [11, 31].

Table 2.2: Details of papers included in the systematic review and an overview of the communication skills domains reported in each.

Author , year	n of students	Length of stations (minutes)	Validity , Reliability (V=validity , R=reliability)	Measurement Instruments	Study Year	Examiners domains	SP raters domains	Professional boards or organizations
Al-Naami, 2008	64	5	V , R	n/a	Final year surgical clerkship	Generic CS, history taking, focused history taking		
Bergus et al., 2009	51	15	R	n/a	3 rd	Generic CS, Physician-patient communication		
Blue et al., 1998	89	n/a	R	n/a	3 rd	History taking, focused history taking		
Blue et al., 2000	476	8	V , R	SPSQ	3 rd	Focused history taking, interview, negotiating plan/treatment, information giving, empathy, emotion/respond of emotion	Generic CS , interpersonal skills	
Blue et al., 2000	205	8	V , R	WHACS, checklist	3 rd	Focused history taking, interview, empathy		
Boehlecke et al., 1996	155	5	R	Checklist	2 nd	Focused history taking		
Bosse et al., 2012	103	n/a	R	Calgary-Cambridge	5 th	Physician-patient communication, history taking, counseling, consultation, health beliefs, interpersonal skills		
Cave et al., 2007	396	5	V	Checklists	3 rd	Generic CS, introduction, history taking, focused history taking, negotiating plan/treatment, information giving		

Author , year	n of students	Length of stations (minutes)	Validity , Reliability (V=validity , R=reliability)	Measurement Instruments	Study Year	Examiners domains	SP raters domains	Professional boards or organizations
Chesser et al., 2004	192	5	V , R	n/a	Penultimate undergraduate year	Generic CS, history taking, focused history taking		
Chessman et al., 2003	127	8 and 15	V	SPSQ , PPQ, ABIM PSQ	3 rd , 4 th	Generic CS	Generic CS, interpersonal skills	ABIM
Harasym et al., 2008	190	n/a	R	Calgary-Cambridge	Family Medicine rotation	Focused history taking		
Ho et al., 2010	57	n/a	V	Checklists	5 th		Generic CS	
Hodges and McIlroy, 2003	57	10	V , R	Checklist , global rating scale	3 rd , 4 th	Non-verbal communication, empathy		
Huang et al., 2010	256	10 and 20	R	n/a	7 th	Generic CS, history taking,		
Humphris, 2002	383	5 and 10	V , R	LCAS, GSPRS	1 st , 2 nd	Introduction, non-verbal communication, empathy	Generic CS	
Jacobs et al., 2004	356	5	V	Maas-R	5 th	Introduction, history taking, focused history taking, negotiating plan/treatment, information giving, interpersonal skills, empathy		

Author , year	n of students	Length of stations (minutes)	Validity , Reliability (V=validity , R=reliability)	Measurement Instruments	Study Year	Examiners domains	SP raters domains	Professional boards or organizations
Kaul et al., 2012	279	n/a	R	Checklists	3 rd		Generic CS, history taking	ACGME
Mazor et al., 2005	100	n/a	V , R	Checklists	3 rd	Negotiating plan/treatment, information giving, health beliefs, empathy	Generic CS, interpersonal skills, health beliefs, empathy	
McLay et al., 2002	82	n/a	V , R	n/a	3 rd	Interview	Generic CS, interpersonal skills	NBME
Park et al., 2004	286	15	R	Checklists, global rating scale	3 rd	History taking, focused history taking, interview	Generic CS, interpersonal skills	NBME
Regehr et al., 1999	161	n/a	V , R	n/a	2 nd , 3 rd	Generic CS, history taking		
Robins et al., 2001	71	20	R	Checklists	4 th	Cross-cultural communication, health beliefs	Health beliefs	
Rosebraugh et al., 1997	196	8	R	n/a	3 rd	History taking		
Scheffer et al., 2008	113	5 and 8	V	Calgary-Cambridge	2 nd , 3 rd	Generic CS, interview, Non-verbal communication, empathy, micro expression		

Author , year	n of students	Length of stations (minutes)	Validity , Reliability (V=validity , R=reliability)	Measurement Instruments	Study Year	Examiners domains	SP raters domains	Professional boards or organizations
Thistlethwaite , 2002	194	6	R	Checklists	3 rd	History taking, negotiating plan/treatment, information giving		
Troncon, 2006	36	7	V , R	n/a	4 th	Physician-patient communication, history taking		
Van Dalen et al., 2002	161	15	V , R	Maas-global	4 th , 6 th	Generic CS, history taking, focused history taking, negotiating plan/treatment, information giving, breaking bad news		
Verma and Singh, 1994	40	n/a	V	Checklists	Final year	Generic CS, Information giving		
Volkan et al., 2004	169	20	R	Checklists	3 rd	Physician-patient communication, history taking		
Walters et al., 2005	128	6	V , R	n/a	4 th	Generic CS, history taking, phone/electronic communication		
Wass and et al., 2001	214	7	V , R	n/a	Final MBBS examination	Generic CS		
Wass and Jolly , 2001	155	8	V , R	Global rating scale	Final MBBS examination	Generic CS		

Author , year	n of students	Length of stations (minutes)	Validity , Reliability (V=validity , R=reliability)	Measurement Instruments	Study Year	Examiners domains	SP raters domains	Professional boards or organizations
Wilkerson et al., 2010	322	15	V , R	Checklists	Senior medical students	Generic CS, negotiating plan/treatment, counseling, health beliefs, empathy	Interpersonal skills, empathy, health beliefs	
Wong et al., 2007	439	10	V , R	Checklists	Final year	Physician-patient communication, information giving, taking consent, breaking bad news, advising/handle family, interpersonal skills		ACGME

Abbreviations list :

SP : Standardised patient

Generic CS : generic communication skills

SPSQ : Standardised patient satisfaction questionnaire

PPQ : Patient perception questionnaire

PSQ : Patient satisfaction questionnaire

WHACS : a mnemonic, provide a few essential questions on occupational and environmental exposures

LCAS : Liverpool communication skills assessment scale

GSPRS : Global simulated patient rating scale

ABIM : American board of internal medicine

ACGME : Accreditation council for graduate medical education

ACGME : Accreditation council for graduate medical education

Chessman (2003), Park (2004), Wong (2007) Kaul (2012) and McLay in 2002 referred to professional board/licensing bodies in describing the design of their instruments [31, 34, 46, 51, 53]. These included the ACGME (Accreditation Council for Graduate Medical Education), NBME (National Board of Medical Examiner) and ABIM (American Board of Internal Medicine).

Fourteen papers reported the use of checklists or global rating scales [15, 28, 29, 34, 35, 38, 41, 43, 45, 46, 50-52, 55], while global rating scales were described in 3 of the 14 papers by Park, 2004, Wass 2001 and Hodges 2003 respectively [15, 34, 43], in Park's paper students were assessed by SP raters only [34]. However we identified 11 papers not reporting any measurement instruments in terms of a documented list of items described in the paper or an appendix [24, 25, 30, 32, 36, 39, 42, 44, 47, 53, 54]. Based upon the information provided in the title, abstract and content, they were excluded from our search results.

Table 2.3: Communication skills domains agreement between 3 reviewers ICC= Intra Class Correlation coefficient Full agreement between reviewers (ICC = 1) means 100% agreement on items assessed. No agreement (ICC = 0) reviewers don't agree at all on items assessed.

	Examiners domains																				SP raters domains						
	Generic CS	Doctor-patient communication	Introduction	History taking	Focussed history taking	Interview	Focussed interview	Negotiating Plan/Treatment	Taking consent	Information giving	Counseling	Consultation	Breaking bad news	Cross-cultural communication	Health beliefs	Advising/handle family	Interpersonal skills	Non-verbal communication	Empathy	Micro expression	Emotion/respond emotion	Phone/electronic communication	Generic CS	History taking	Interpersonal skills	Empathy	Health beliefs
Percentage Papers	52	34	16	58	41	25	9	30	9	27	14	10	11	7	20	6	17	17	30	6	8	7	31	6	25	14	15
ICC	0.83	0.69	0.86	0.81	0.9	0.82	-0.12	0.85	0.73	0.92	0.74	0.66	0.9	0.9	0.85	1	0.82	0.82	0.93	1	0.82	0.9	0.91	1	0.92	0.75	0.89

2.4.3. Standard Setting

The Maas-Global, the first available standard proven to be valid and reliable, consists of a check-list and a 20-page scoring manual, listing criteria per item [40]. The focus of this instrument is on the communication process, rather than the content, i.e. how questions are asked rather than what is asked [40]. Simone Scheffer validated a Global Rating Scale assessing empathy, degree of coherence in the interview, verbal expression and non-verbal expression [37]. In her study encounters were evaluated using the short version of the Calgary Cambridge Observation Guide. This Guide divides communication in medical settings into two broad categories: (a) interviewing the patient and (b) explanation and planning. Each of the categories has several components. For example, interviewing the patient is further divided into (a) initiating the session, (b) gathering information, (c) building relationship, and (d) explaining and planning [57]. According to the CCOG, this guide can be used as checklist for CS assessment and as feedback tool to the learner although publications on reliability and validity of the CCOG as an assessment tool are lacking. However many of the checklists we found used the CCOG as a kind of standard. The Standardized Patient Satisfaction Questionnaire (SPSQ) scores students in the following performance domains: 1) interviewing skills, 2) negotiating the diagnosis or plan, 3) gathering case-specific content information, 4) responding to the patient's emotions, and 5) student's overall performance. Pearson Product-Moment correlations were calculated for each of these domains [26].

2.4.4. Reviewer Agreement

Agreement between our reviewers, expressed in an Intraclass Correlation Coefficient (ICC) on the CS domains, ranged from -0.12 to 1 and the ICC on all CS domains was 0.81, while total ICC on all marked items was 0.68.

Agreement improved after the reviewers discussed items whereby only two out of three agreed initially. For the purposes of presenting the results, the situation

where reviewers were in full agreement prior to such discussion is termed “initial agreement”. The situation whereby reviewers achieved full agreement after discussing the disagreed item(s) is termed “resolved disagreement”. The comparison between initial agreement (17%) and resolved disagreement (83%) for measurement instruments amongst reviewers is illustrated in Figure 2. For CS domains, initial agreement was 33% and this increased to 67% upon discussion. Meanwhile 'n of student' and 'duration of station' had low percentage of resolved disagreement and 'validity/reliability' were 50%.

Resolved disagreement : Initial agreement

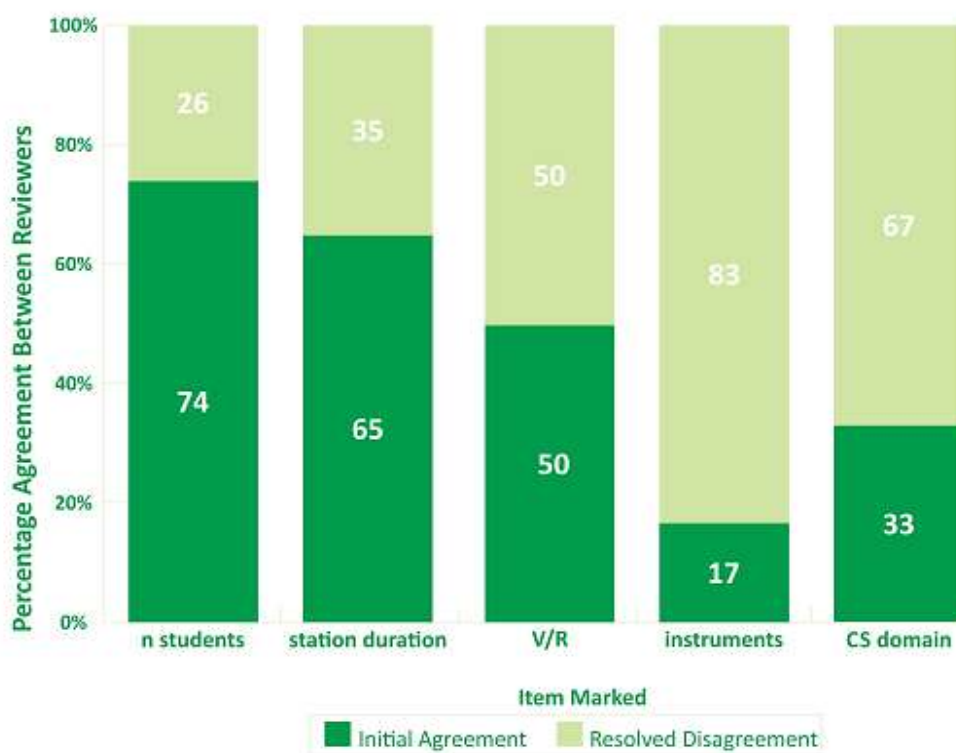


Figure 2.2: Comparison between resolved disagreement and initial agreement amongst reviewers.

* V / R = validity / reliability; CS Domains = Communication Skills domains

2.5. Discussion and Conclusion

2.5.1. Discussion

The most striking finding of our study is a demonstrated absence of consensus in rubrics used to assess CS in undergraduate medical education worldwide. Furthermore, it is apparent that there is a clear absence of consensus between researchers in medical education in their interpretation of terminology and in their determination of performance standards in the assessment of CS in different settings. The OSCE is widely utilized to assess CS at undergraduate and postgraduate levels [55]. It is likely that significant heterogeneity exists in teaching and assessment of CS across different institutions as well as across different years of the curriculum within the same institution. One possible explanation is that those who use established local instruments do not frequently publish their adaptation or validation in their circumstances. It is perhaps not surprising that there are differences between tools designed to assess different CS. Similarly, it is not surprising that there are differences in tools designed to assess measurement of blood pressure and those used to assess performance of basic life support. However, the demonstrated lack of agreement and transparency in the use of terminology and lack of published reliability and validity of any type of OSCE station is of concern. This absence of standardization of assessment rubrics in undergraduate medical education precludes the comparison of outcomes across assessment settings. This study highlights the absence of an agreed gold standard for the assessment of CS of undergraduate medical students. This finding is of particular note in the context of the existence of the first reliable and valid measurement tool, known as the MAAS Global [58].

We identified only 9 papers (27%) which did not report station duration and 4 studies (11%) which reported two different durations (a short case and a long case approach). These results align with the findings of Patricio et al (2013) who concluded that only 30% of papers reported station duration [59]. This finding is

of concern because station duration is known to be one determinant of station reliability. Effective assessment of CS requires enough time to adequately cover the objective of the communication. Thus, we recommend that future research pertaining to the assessment of CS should always report station duration and the objective of the communication so that research findings can be more easily compared and synthesized.

The majority of included studies did not clearly report measurement instruments and the underlying construct of various CS domains was unclear in 19 out of 23 (83%) papers (Figure 2). It is apparent from Figure 2 that the included research on CS assessment can very easily be misinterpreted, even by expert reviewers. For instance, reviewer disagreement upon the number of student participants, an item that should be very clear, was only resolved after two meetings.

Difficulty arose during the reviewer analysis of the papers with respect to the CS domains that were measured and the interpretation of the terminology used to describe such domains. The ICC for all domains described in the papers included in the review was 0.81. With respect to the domain of 'focused interview,' we didn't achieve agreement (ICC -0.12). This finding is explained by the absence of any description of the 'focused interview' domain in all of those papers [26, 27, 33, 34, 45, 49]. Full agreement was reached in only three of the included papers. Each of these papers described only one domain of CS. The crucial omission of clear descriptions of CS domains was previously described by Boon and Steward (1998), Beck et al (2002) and Cegala and Broz (2002) [4, 19, 60, 61]. A clear description of the object or underlying concepts in the relationship with empirical indicators is the single most important requirement for assessment [62-64]. If the concept that is to be assessed is not clearly defined and clear indicators are not included, then it cannot be adequately measured. We suggest that educational decisions drawn from flawed measures are unreliable.

The majority of included studies focused upon physician-patient interaction. However, in reality, physicians must also be able to communicate effectively with other physicians, with nurses and with other stakeholders [65-70]. Our review demonstrates a notable paucity of published research in this field. Furthermore, we identified only one study which explored the assessment of the use of phone/electronic communication at undergraduate level [42]. Other forms of communication include interpersonal skills, non-verbal communication, micro expression and empathy. We identified eight papers which studied the assessment of empathy, suggesting that this domain of CS is an apparent priority for researchers [11, 15, 26, 27, 33, 37, 45, 52].

Reliability of results and validity of CS are essential to the assessment of student competence [64]. There are other opinions according to which reliability of results is a prerequisite of validity while others mention that reliability of results is necessary but not sufficient for the sole support of validity [71]. We found that 16 (47%) of the reviewed papers reported both reliability of results and validity (internal consistency), whilst the remainder reported only one of these two measures. It was notable that the majority of papers did not refer to a recognized Gold standard with a view to improving the construct validity of each assessment form used.

In contrast, the University of Maastricht developed a unique and validated instrument, currently known as the MAAS-Global, which was first reported back in 1990 [58, 72]. This instrument is being used, in real-time and in recorded sessions, to assess students, physicians and/or nurses at only a small number of institutions in different countries. However, it is important to recognize that it is not being more widely adopted as a gold standard [73-75]. It is apparent from this review that the majority of CS assessment is based upon the individual development of unique measurement instruments which are used only at local level. On the contrary, we only found three instruments the MAAS-Global

(including MAAS-R), the SPSQ, and the CCOG which were reported in more than one study. The CCOG is actually a guideline which was not designed to be used as a validated assessment instrument. As demonstrated in previous reviews published in 1998 and 2002, the present review again identified a failure to adopt existing validated instruments [4, 19]. Inability to reproduce results across assessments precludes meaningful interpretation of results [76].

Whilst some time has passed since the aforementioned reviews in 1998 and 2002 were carried out, the two main problems of how to define an appropriate learning outcome of a specific CS domain and an appropriate method of measurement still exist in 2015. Rather than repeatedly creating new assessment forms, researchers and educators need to work together in order to agree upon the definitions of learning outcomes and CS domains, so that gold standard CS measurement instruments can be developed. Bloch and Norman, 2012, doubt whether competence can be measured with a single scale (i.e. one measurement instrument for all CS), as opposed to a unique scale (i.e. a specific measurement for each specific domain of CS) for different specialties and different practice conditions [76]. While 're-inventing the wheel' is not necessary, any effort to incorporate or modify existing instruments in order to fit into different specialties and practice conditions will be valuable for future development of undergraduate medical education. The Step 2 CS is a high stakes CS assessment tool. The Clinical Skills Review (CSR) is an Interactive Internet based preparatory site for the United States Medical Licensing Examination (USMLE) Step 2 Clinical Skills (CS) live exam. CSR offers a specialized learning environment that is aligned with the rules and regulations set forth by the official exam provider. We agree that there should be alignment between undergraduate CS training with the expected learning outcomes and assessment goals of these high stake licensing exams. In short, we suggest that standardization (i.e. uniform use of valid, reliable and aligned CS measurement instruments) and alignment of undergraduate and postgraduate communication skills training is necessary in

order to sufficiently meet the requirements of professional practice. Whilst global standardization might be very challenging, we wish to highlight the importance of standardization and appropriate use of statistics as a prerequisite for student outcome comparison between and across local and national settings [77].

Limitations of this systematic review include the exclusion of studies published in languages other than English and those not pertaining to undergraduate medical students, thus it may not be appropriate to generalize results to assessment in other student populations and settings. In retrospect, we did not adequately take into consideration the importance of aligning postgraduate and undergraduate training and assessment of CS and the use of frameworks like CanMeds and others to highlight appropriate 'top down' alignment of CS training. Furthermore, despite rigorous research methods, incomplete retrieval of published literature is possible. Despite CS being an important competence for students to master, the assessment of CS continues to be a challenging endeavor. In the US it is now mandated that medical students and residents have CS training and the variety and variances in this training are as numerous as there are programs. Internationally, there are still medical schools that have not incorporated CS training in a formal way. Clearly there is work to be done and reviewing what we know to work well is important, starting with a clear description of the underlying concepts.

2.5.2. Conclusion

We demonstrate a clear absence of consensus between researchers in their interpretation and definition of domains of CS. Included papers generally failed to satisfactorily identify the underlying constructs and learning outcomes that were being assessed. Terminology was not uniformly employed across included papers.

Furthermore, there was poor consistency with respect to the use of Likert scales and global ratings scales, despite this issue having been previously identified [19]. A valid and reliable measurement instrument, such as the MAAS-Global (<http://bit.ly/1xQXAnS>), is not universally accepted and this paper promotes calibration of communication skills using this valid and reliable standard.

2.6. Practice Implications

Future research should focus upon the comparison of the clinical skills stations in our and other institutions using the Maas-Global as a standard to calibrate existing CS items in each of the assessment forms, so that measures of CS become interchangeable and comparable within and between institutions. We suggest that such calibration could be based upon the Maas-Global.

2.7. References for the Study

- [1] Stewart MA. Effective physician-patient communication and health outcomes: a review. *Can Med Ass J* 1995;152(9):1423-33.
- [2] Ong LM, de Haes JC, Hoos AM, Lammes FB. Doctor-patient communication: a review of the literature. *Soc Sci Med* 1995;40(7):903-18.
- [3] Phillips C. Communication: the first tool in risk management for long-term care. *J Am Med Dir Assoc*. 2004; 5(2):123-6.
- [4] Beck RS, Daughtridge R, Sloane PD. Physician-patient communication in the primary care office: a systematic review. *J Am Board Fam Pract* 2002;15(1):25-38.
- [5] Levinson W. Physician-patient communication. A key to malpractice prevention. *J Amer Med Assoc* 1994; 272(20):1619-20.
- [6] Maguire P, Pitceathly C. Key communication skills and how to acquire them. *Brit Med J*. 2002; 28;325(7366):697-700.
- [7] Maguire P, Pitceathly C. Managing the difficult consultation. *Clin Med* 2003; 3(6):532-7.
- [8] Noble LM, Kubacki A, Martin J, Lloyd M. The effect of professional skills training on patient-centredness and confidence in communicating with patients. *Med Educ* 2007; 41(5):432-40.
- [9] Bowleg L, Valera P, Teti M, Tschann JM. Silences, gestures, and words: nonverbal and verbal communication about HIV/AIDS and condom use in black heterosexual relationships. *Health Commun* 2010; 25(1):80-90.
- [10] Yedidia MJ, Gillespie CC, Kachur E, Schwartz MD, Ockene J, Chepaitis AE, et al. Effect of communications training on medical student performance. *J Amer*

Med Assoc 2003; 290(9):1157-65.

[11] Humphris GM. Communication skills knowledge, understanding and OSCE performance in medical trainees: a multivariate prospective study using structural equation modelling. *Med Educ* 2002; 36(9):842-52.

[12] Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J* 1975; 22;1(5955):447-51.

[13] Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. *Fam Med* 2008; 40(8):574-8.

[14] Hodges B, Turnbull J, Cohen R, Bienenstock A, Norman G. Evaluating communication skills in the OSCE format: reliability and generalizability. *Med Educ* 1996; 30(1):38-43.

[15] Hodges B, McIlroy JH. Analytic global OSCE ratings are sensitive to level of training. *Med Educ*. 2003; 37(11):1012-6.

[16] Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. *Acad Med* 2003; 78(10 Suppl):S33-5.

[17] Khattab AD, Rawlings B. Assessing nurse practitioner students using a modified objective structured clinical examination (OSCE). *Nurse Educ Today* 2001; 21(7):541-50.

[18] Khattab AD, Rawlings B. Use of a modified OSCE to assess nurse practitioner students. *Br J Nurs* 2008; 17(12):754-9.

[19] Boon H, Stewart M. Patient-physician communication assessment instruments: 1986 to 1996 in review. *Patient Educ Couns* 1998; 35(3):161-76.

[20] Barry M, Bradshaw C, Noonan M. Improving the content and face validity of OSCE assessment marking criteria on an undergraduate midwifery programme: a quality initiative. *Nurse Educ Pract* 2013; 13(5):477-80.

[21] Moineau G, Power B, Pion AM, Wood TJ, Humphrey-Murto S. Comparison of student examiner to faculty examiner scoring and feedback in an OSCE. *Med Educ* 2011; 45(2):183-91.

[22] Chang A, Boscardin C, Chou CL, Loeser H, Hauer KE. Predicting failing performance on a standardized patient clinical performance examination: the importance of communication and professionalism skills deficits. *Acad Med* 2009; 84(10 Suppl):S101-4.

[23] Schirmer JM, Mauksch L, Lang F, Marvel MK, Zoppi K, Epstein RM, et al. Assessing communication competence: a review of current tools. *Fam Med* 2005; 37(3):184-92.

[57] Anonymous. Calgary-Cambridge Observation Guide; Appendix A2002; (Basics in Medical Education): Available from: http://www.worldscientific.com/doi/pdf/10.1142/9789812795472_bmatter.

[58] van Es JM, Schrijver CJ, Oberink RH, Visser MR. Two-dimensional structure of the MAAS-Global rating list for consultation skills of doctors. *Med Teach* 2012; 34(12):e794-9.

- [59] Patricio MF, Juliao M, Fareleira F, Carneiro AV. Is the OSCE a feasible tool to assess competencies in undergraduate medical education? *Med Teach* 2013; 35(6):503-14.
- [60] Cegala DJ, Gade C, Lenzmeier Broz S, McClure L. Physicians' and patients' perceptions of patients' communication competence in a primary care medical interview. *Health Commun* 2004; 16(3):289-304.
- [61] Cegala DJ, Lenzmeier Broz S. Physician communication skills training: a review of theoretical backgrounds, objectives and skills. *Med Educ* 2002 Nov; 36(11):1004-16.
- [62] Beckman TJ, Cook DA. Educational epidemiology. *J Amer Med Assoc* 2004; 22;292(24):2969; author reply 70-1.
- [63] Beckman TJ, Cook DA, Mandrekar JN. What is the validity evidence for assessments of clinical teaching? *J Gen Intern Med* 2005; 20(12):1159-64.
- [64] Beckman TJ, Ghosh AK, Cook DA, Erwin PJ, Mandrekar JN. How reliable are assessments of clinical teaching? A review of the published instruments. *J Gen Intern Med* 2004; 19(9):971-7.
- [65] Fallowfield L, Jenkins V. Effective communication skills are the key to good cancer care. *Eur J Cancer* 1999; 35(11):1592-7.
- [66] Fallowfield L, Jenkins V. Acronymic trials: the good, the bad, and the coercive. *Lancet* 2002; 23;360(9346):1622.
- [67] Fallowfield L, Jenkins V. Communicating sad, bad, and difficult news in medicine. *Lancet* 2004; 363(9405):312-9.
- [68] Fallowfield L, Jenkins V. Current concepts of communication skills training in oncology. *Recent Results Cancer Res.* 2006;168:105-12.
- [69] Vazirani S, Hays RD, Shapiro MF, Cowan M. Effect of a multidisciplinary intervention on communication and collaboration among physicians and nurses. *Am J Crit Care* 2005 ;14(1):71-7.
- [70] Beuscart-Zephir MC, Pelayo S, Anceaux F, Meaux JJ, Degroisse M, Degoulet P. Impact of CPOE on doctor-nurse cooperation for the medication ordering and administration process. *Int J Med Inform* 2005;74(7-8):629-41.
- [71] Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med* 2006; 119(2):166 e7-16.
- [72] Kraan HF, Crijnen AA, de Vries MW, Zuidweg J, Imbos T, Van der Vleuten CP. To what extent are medical interviewing skills teachable? *Med Teach* 1990;12(3-4):315-28.
- [73] Essers G, Kramer A, Andriess B, van Weel C, van der Vleuten C, van Dulmen S. Context factors in general practitioner-patient encounters and their impact on assessing communication skills--an exploratory study. *BMC Fam Pract* 2013;14:65.
- [74] Hobma S, Ram P, Muijtjens A, van der Vleuten C, Grol R. Effective improvement of doctor-patient communication: a randomised controlled trial. *Br J Gen Pract* 2006; 56(529):580-6.

- [75] Reinders ME, Blankenstein AH, van Marwijk HW, Knol DL, Ram P, van der Horst HE, et al. Reliability of consultation skills assessments using standardised versus real patients. *Med Educ* 2011; 45(6):578-84.
- [76] Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach* 2012; 34(11):960-92.
- [77] step2cs.net Copyright© 2009-2015 Clinical Skills Review, LLC. All Rights Reserved. Legal@Step2CS.net

2.8. References for the Review

- [24] Al-Naami MY. Reliability, validity, and feasibility of the Objective Structured Clinical Examination in assessing clinical skills of final year surgical clerkship. *Saudi Med J* 2008; 29(12):1802-7.
- [25] Bergus GR, Woodhead JC, Kreiter CD. Trained lay observers can reliably assess medical students' communication skills. *Med Educ* 2009; 43(7):688-94.
- [26] Blue AV, Chessman AW, Gilbert GE, Mainous AG, 3rd. Responding to patients' emotions: important for standardized patient satisfaction. *Fam Med* 2000; 32(5):326-30.
- [27] Blue AV, Chessman AW, Gilbert GE, Schuman SH, Mainous AG. Medical students' abilities to take an occupational history: use of the WHACS mnemonic. *J Occup Environ Med* 2000; 42(11):1050-3.
- [28] Boehlecke B, Sperber AD, Kowlowitz V, Becker M, Contreras A, McGaghie WC. Smoking history-taking skills: a simple guide to teach medical students. *Med Educ* 1996; 30(4):283-9.
- [29] Cave J, Washer P, Sampson P, Griffin M, Noble L. Explicitly linking teaching and assessment of communication skills. *Med Teach* 2007; 29(4):317-22.
- [30] Chesser AM, Laing MR, Miedzybrodzka ZH, Brittenden J, Heys SD. Factor analysis can be a useful standard setting tool in a high stakes OSCE assessment. *Med Educ* 2004; 38(8):825-31.
- [31] Chessman AW, Blue AV, Gilbert GE, Carey M, Mainous AG, 3rd. Assessing students' communication and interpersonal skills across evaluation settings. *Fam Med* 2003; 35(9):643-8.
- [32] Huang CC, Chan CY, Wu CL, Chen YL, Yang HW, Chen CH, et al. Assessment of clinical competence of medical students using the objective structured clinical examination: first 2 years' experience in Taipei Veterans General Hospital. *J Chin Med Assoc* 2010; 73(11):589-95.
- [33] Jacobs JC, Denessen E, Postma CT. The structure of medical competence and results of an OSCE. *Neth J Med* 2004; 62(10):397-403.
- [34] Park RS, Chibnall JT, Blaskiewicz RJ, Furman GE, Powell JK, Mohr CJ. Construct validity of an objective structured clinical examination (OSCE) in psychiatry: associations with the clinical skills examination and other indicators. *Acad Psychiatry* 2004; 28(2):122-8.
- [35] Robins LS, White CB, Alexander GL, Gruppen LD, Grum CM. Assessing

medical students' awareness of and sensitivity to diverse health beliefs using a standardized patient station. *Acad Med* 2001; 76(1):76-80.

[36] Rosebraugh CJ, Speer AJ, Solomon DJ, Szauter KE, Ainsworth MA, Holden MD, et al. Setting standards and defining quality of performance in the validation of a standardized-patient examination format. *Acad Med* 1997; 72(11):1012-4.

[37] Scheffer S, Muehlinghaus I, Froehmel A, Ortwein H. Assessing students' communication skills: validation of a global rating. *Adv Health Sci Educ Theory Pract* 2008;13(5):583-92.

[38] Thistlethwaite JE. Developing an OSCE station to assess the ability of medical students to share information and decisions with patients: issues relating to interrater reliability and the use of simulated patients. *Educ Health (Abingdon)* 2002;15(2):170-9.

[39] Troncon LE. Significance of experts' overall ratings for medical student competence in relation to history-taking. *Sao Paulo Med J* 2006; 124(2):101-4.

[40] van Dalen J, Kerkhofs E, van Knippenberg-Van Den Berg BW, van Den Hout HA, Scherpbier AJ, van der Vleuten CP. Longitudinal and concentrated communication skills programmes: two dutch medical schools compared. *Adv Health Sci Educ Theory Pract* 2002;7(1):29-40.

[41] Volkan K, Simon SR, Baker H, Todres ID. Psychometric structure of a comprehensive objective structured clinical examination: a factor analytic approach. *Adv Health Sci Educ Theory Pract* 2004;9(2):83-92.

[42] Walters K, Osborn D, Raven P. The development, validity and reliability of a multimodality objective structured clinical examination in psychiatry. *Med Educ* 2005;39(3):292-8.

[43] Wass V, Jolly B. Does observation add to the validity of the long case? *Med Educ* 2001;35(8):729-34.

[44] Wass V, McGibbon D, Van der Vleuten C. Composite undergraduate clinical examinations: how should the components be combined to maximize reliability? *Med Educ* 2001;35(4):326-30.

[45] Wilkerson L, Fung CC, May W, Elliott D. Assessing patient-centered care: one approach to health disparities education. *J Gen Intern Med* 2010; 25 Suppl 2:S86-90.

[46] Wong ML, Fones CS, Aw M, Tan CH, Low PS, Amin Z, et al. Should non-expert clinician examiners be used in objective structured assessment of communication skills among final year medical undergraduates? *Med Teach* 2007; 29(9):927-32.

[47] Blue AV, Stratton TD, Plymale M, DeGnore LT, Schwartz RW, Sloan DA. The effectiveness of the structured clinical instruction module. *Am J Surg* 1998; 176(1):67-70.

[48] Bosse HM, Schultz JH, Nickel M, Lutz T, Moltner A, Junger J, et al. The effect of using standardized patients or peer role play on ratings of undergraduate communication training: a randomized controlled trial. *Patient Educ Couns* 2012;87(3):300-6.

- [49] Harasym PH, Woloschuk W, Cunning L. Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs. *Adv Health Sci Educ Theory Pract* 2008;13(5):617-32.
- [50] Ho MJ, Yao G, Lee KL, Hwang TJ, Beach MC. Long-term effectiveness of patient-centered training in cultural competence: what is retained? What is lost? *Acad Med* 2010;85(4):660-4.
- [51] Kaul P, Barley G, Guiton G. Medical student performance on an adolescent medicine examination. *J Adolesc Health* 2012;51(3):299-301.
- [52] Mazor KM, Ockene JK, Rogers HJ, Carlin MM, Quirk ME. The relationship between checklist scores on a communication OSCE and analogue patients' perceptions of communication. *Adv Health Sci Educ Theory Pract*. 2005;10(1):37-51.
- [53] McLay RN, Rodenhauser P, Anderson DS, Stanton ML, Markert RJ. Simulating a full-length psychiatric interview with a complex patient: an OSCE for medical students. *Acad Psychiatry* 2002;26(3):162-7.
- [54] Regehr G, Freeman R, Hodges B, Russell L. Assessing the generalizability of OSCE measures across content domains. *Acad Med* 1999;74(12):1320-2.
- [55] Verma M, Singh T. Communication skills in clinical practice fad or necessity? *Indian Pediatr* 1994;31(2):237-8.
- [56] Schuman SH, Simpson WM, Jr. WHACS your patients. *J Occup Environ Med* 1999;41(10):829.

Chapter 3. Calibration of Communication Skills Items in OSCE Checklists According to the MAAS-Global

This chapter was published as a paper in the Patient Education and Counseling, 2015, <http://dx.doi.org/10.1016/j.pec.2015.08.001>

3.1. Abstract

Background. Communication skills (CS) are commonly assessed using 'communication items' in Objective Structured Clinical Examination (OSCE) station checklists. Our aim is to calibrate the communication component of OSCE station checklists according to the MAAS-Global which is a valid and reliable standard to assess CS in undergraduate medical education.

Method. Three raters independently compared 280 checklists from 4 disciplines contributing to the undergraduate year 4 OSCE against the 17 items of the MAAS-Global standard. G-theory was used to analyze the reliability of this calibration procedure.

Results. G-Kappa was 0.8. For two raters G-Kappa is 0.72 and it fell to 0.57 for one rater. 46% of the checklist items corresponded to section three of the MAAS-Global (i.e. medical content of the consultation), whilst 12% corresponded to section two (i.e. general CS), and 8.2% to section one (i.e. CS for each separate phase of the consultation). 34% of the items were not considered to be CS.

Conclusion. A G-Kappa of 0.8 confirms a reliable and valid procedure for calibrating OSCE CS checklist items using the MAAS-Global. We strongly suggest that such a procedure is more widely employed to arrive at a stable (valid and reliable) judgment of the communication component in existing checklists for medical students' communication behaviors.

Practice Implications. It is possible to measure the 'true' caliber of CS in OSCE stations. Students' results are thereby comparable between and across stations,

students and institutions. A reliable calibration procedure requires only two raters.

3.2. Background

The Objective Structured Clinical Examination (OSCE) is commonly used to assess the communication skills (CS) of undergraduate medical students. Curriculum design frequently starts with blue printing learning outcomes, teaching and assessment methods according to the Best Evidence Medical Education guidelines [BEME] [1]. The lack of clear descriptions of CS domains in OSCE's has previously been identified by Boon and Steward (1998), Beck et al (2002) and Cegala and Broz (2002) [2–4]. We highlighted the existence of 27 domains of CS and a lack of clarity with respect to which of these domains are thought in medical curricula [5]. Furthermore, there is no agreed universally accepted standard for the assessment of the CS of undergraduate medical students [5]. Key concepts (e.g. specificity, blue printing, feasibility and global rating versus checklist rating) are not often being explicitly addressed as is suggested by professional bodies like the European and American Association for Communication in Healthcare [6,7]. This absence of blue printing and standardization precludes the comparison of outcomes across assessment settings. Calibration in terms of validation and standardisation of a measurement tool used for assessment purposes is crucial.

Reproducibility of results and validity of CS assessments are essential to the measurement of student competence [8]. Approximately half of the published research papers reported reproducibility of results and validity (internal consistency) [5,9]. Internationally at least three frameworks for the analysis of doctor-patient communication are acknowledged and used in a global context: the Calgary-Cambridge Observation Guides, Roter's Interaction Analysis System (RIAS) and the MAAS-Global [5,10,11]. Whilst these are useful tools for informing CS education strategy, they are not necessarily valid for assessment of medical

student CS. The developers of Calgary-Cambridge Observation guides, for example, never intended the guides to be used as a checklist of observable skills informative teaching [12]. They were not designed to be measuring instruments. Experts in the field of CS education, including the chief developer of the Calgary-Cambridge guide, have previously expressed caution with respect to its misuse [12,13]. There is no generally accepted measurement instrument (i.e. agreed upon by all researchers) for the assessment of CS in undergraduate medical students [4,5]. We chose to explore the external validity of the assessment instruments used in our medical school using the MAAS Global because of its use in previous undergraduate comparative studies [11].

The purpose of this study is to calibrate existing CS assessment forms being used in our medical school. We compare the estimates of three raters externally validating the CS items contained in our existing forms to see whether they match with the MAAS-Global.

3.3. Methods

3.3.1. Context of the Study

We evaluated all station checklists (measurement instruments) used by the Disciplines of Obstetrics & Gynaecology, Paediatrics, Psychiatry, and General Practice in year 4 of the undergraduate medical programme at the National University of Ireland in Galway, Ireland. Year 4 is the penultimate year of the undergraduate medical programme.

3.3.2. Description of the OSCE

Four disciplines – Disciplines of Obstetrics & Gynaecology, Paediatrics, Psychiatry, and General Practice – contribute to the year 4 OSCE using their own discipline-specific stations. Station forms were made available in the station bank of the OSCE Management Information System (OMIS), as the OSCE was planned and executed [14]. The specific order of stations in an OSCE examination which

allows students to follow through the consecutive station examinations is called a circuit [15]. Each discipline contributing to the year 4 OSCE uses different circuit settings, such as number of stations, sequence of stations, and/or scoring rubrics (assessment forms). The duration of all stations is set to 5 minutes with 1 minute in between stations and 1 minute reading time prior to the start of each station.

The data from four academic terms – 2009/2010, 2010/2011, 2011/2012, and 2012/2013 – (number of students total = 454 i.e. 115 (2009/10); 118 (2010/11); 123 (2011/12) and 140 (2012/13) respectively) were retrospectively analyzed. In total, 250 assessment forms used in 27 OSCE circuits (Table 3.1) were analyzed. Further details of contributions from each discipline are presented in Table 3.1.

Table 3.1: Summary of OSCE's circuits, stations, and checklist's items in each disciplines in 3 academic terms (2010/2011, 2011/2012, and 2012/2013).

Disciplines	Academic Year	OSCE circuits	Stations per circuit	Checklist' items per circuit
Discipline of Obstetrics & Gynaecology	2010 – 2011	2	10	60
	2011 – 2012	2	10	51
	2012 – 2013	2	10	96
Discipline of Paediatrics	2009 – 2010	1	6	122
	2010 – 2011	2	7 & 9	210
	2011 – 2012	2	10	220
Discipline of General Practice	2012 – 2013	2	10	229
	2009 – 2010	1	8	117
	2010 – 2011	2	10	349
Discipline of Psychiatry	2011– 2012	2	10	231
	2012 – 2013	2	10	202
	2009 – 2010	1	4	75
Discipline of Psychiatry	2010 – 2011	2	8	170
	2011 – 2012	2	10	250
	2012 – 2013	2	10	195
	Total			2577

3.3.3. Calibration Checklists

In this study, the term ‘calibration’ is used to rate how close the items in the stations’ checklist(s) fit the MAAS-Global standard. The rationale for choosing the MAAS-Global is that it was developed as a measuring tool with known validity and reliability [16]. Furthermore, the MAAS-Global is designed as a generic instrument to rate physicians' CS and has been previously used to compare undergraduate medical students [11]. The MAAS-Global consists of 17 items divided into 3 sections. Seven items in section 1 refer to appropriate skills in the specific phases of clinical consultations. Items are related to introduction, follow-up consultation, a request for help, physical examination, diagnosis, management, and evaluation of the consultation. These items are a reflection of the logical order of consultation phases. Section 2 focuses on general CS which occur throughout the consultation, consisting of 6 items. Those items are: exploration, emotions, information giving, summarizations, structuring, and empathy. Section 3 is intended to examine the mastery of the medical content during medical consultation. This section consists of 4 items: history taking, physical examination, diagnosis, and management which represent phases of the consultation (see appendix).

We used the MAAS-Global rating list as the independent standard for comparison of each individual item on each checklist used within the year 4 OSCE. The authors created a manual for calibration which was called MAAS-Global Calibration Checklist (MGCC). The manual consists of 3 parts. The first part is an explanatory part on how to rate the station's checklists according to the MAAS-Global. The second part describes the definition of the concept of the MAAS-Global. Finally, the third part is a detailed explanation of each of the items of MAAS-Global. All parts of the manual, except the explanatory part, are based upon the MAAS-Global 2000 Manual (See supporting document entitled “MAAS-Global Check-lists Calibration Manual”).

3.3.4. Choice of Statistical Approach

In classical test theory, consistency in an assessment procedure is usually expressed as inter-observer, intra-observer and test-retest reliability and intraclass correlation coefficients. These coefficients are not measures of quantitative change [17]. The results of reliability studies are specific to the examiner(s) involved in each specific study and are not generalisable to other examiners and assessment settings. In a classical psychometric approach error is calculated as $1 - R$. For example, in a case where inter-observer, intra-observer and test-retest reliability are considered to be good or excellent, with an R of 0.8, there remains a 20% ($1 - 0.8$) “error” around the observed score. In a generalizability study, multiple variance components (i.e. sources of variation such as disciplines, examiners, and station forms and all their interactions) are estimated [18,19]. Classical test theory only recognizes two types of variances: true variance and error variance [20,21]. Whereas in, a Generalizability Theory study, analysis will more appropriately show the contribution of each of the potential sources of error variance to the total error [22]. The Generalizability Theory (G-theory) analysis is complementary to the classical psychometric theory and consists of a Generalizability study (G-study) and a Decision study (D-study). The former identifies the primary sources of variation and their interactions that contribute to the total error variance of a measurement procedure (i.e. the measurement design), whereas the Decision study incorporates the impact of the error variation on the decision to be taken depending on the chosen measurement design regarding passing or failing students in a reliable manner [19,23]. The D-study also expresses measures of change in the unit of the measurement tool employed. We chose to employ G-theory analysis for the present study. Furthermore, whilst using classical psychometric analysis (e.g. Kappa statistics) would also help to identify variation, such analysis would not, in our opinion, provide insight in to the sources of identified variation.

3.3.5. Procedure

Three raters participated in the calibration of the station forms. At first, they met for instructions from the first author about the procedure, (i.e. the method of calibration). The three raters were trained in multiple meetings to mark each item on each station checklist in accordance with the MAAS-Global criteria. To validate the raters' interpretation of the MGCC manual, samples of station checklists (n=6) from each discipline were rated by these three raters independently and discussed in a second meeting (Figure 3.1). The upper portion of Figure 3.1 provides an example of how each individual rater matched OSCE checklist items with the items of the MAAS-Global.

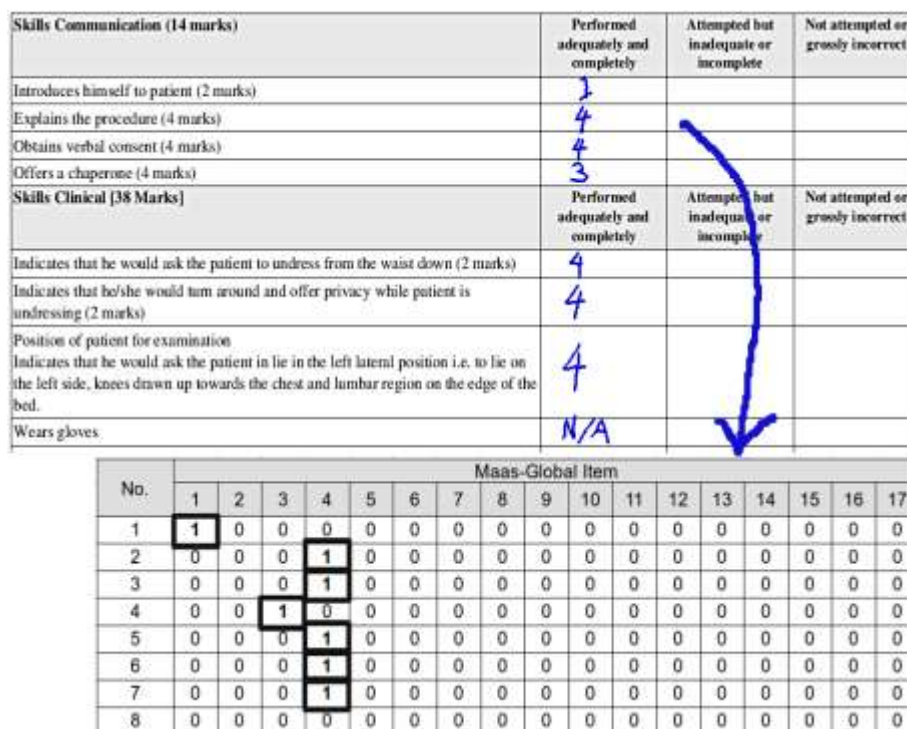


Figure 3.1: Illustration on how raters calibrate station's checklists and then transferred into spreadsheet.

The second meeting among raters was conducted to discuss the result of the rated sample checklists. Discrepancies were discussed until consensus was reached. Each rater received all year 4 station checklists (n=250) from all 4

participating disciplines. Raters independently scored all checklist items according to the 17 items of the MAAS-Global. In the case whereby an item did not match with any of the MAAS-Global items, then this item was assigned the term “not applicable” or “N/A”.

3.3.6. Analysis

Raters matched checklist items with the appropriate MAAS-Global items and noted the MAAS-Global item number. All results were transferred into a spreadsheet containing the 17 nominal (MAAS-Global) items classified as either 'zero' or 'one'. When the checklist items were rated as 'not applicable' (N/A), all columns were filled with zeros (Figure 3.1). The lower portion of Figure 3.1 provides an example of binary translation of the rater’s assessment forms.

The data was analyzed as a 4-facet Generalizability Theory study, with Facet 1 being “Checklists” (nested within disciplines), Facet 2 being “Disciplines”, Facet 3 being “The MAAS-Global Items” and Facet 4 being “Raters”. All facets are ‘random facets’, with the exception of The MAAS-Global Items, which was ‘fixed facet’. The software package “EduG” (Version 6.1-e) was used to analyze the reliability of the calibration process according [18]. G-Kappa is the term used in Generalizability theory to analyze binary data [23]. The accepted G-Kappa value for precision of measurement is 0.80 [18].

3.4. Results

3.4.1. Station Checklists

Descriptive data pertaining to the included OSCEs are presented in Table 1. For logistical reasons, particularly with regard to assessing a large number of students and a large number of competencies, the OSCE in year 4 is delivered bi-annually, with one circuit/round of students going through the OSCE in February and another circuit of students going through a similar OSCE in April. In the academic year 2009-2010, electronic assessment using the in-house developed

OSCE Management Information System (OMIS) (Kropmans 2012) was introduced [14]. Data for this academic year is limited to one circuit of students because electronic OMIS was introduced late in that academic year (March 2009). Each circuit contained a total of between 4 and 15 individual stations. The number of participating stations of each contributing discipline varied between 4 and 10. The number of checklist items varied from 1 to 35 per station. The Discipline of Psychiatry had the highest average of checklist items per station compared to the other three disciplines, of which Obstetrics & Gynaecology had the lowest amount of checklist items. While the Year 4 OSCE was delivered a number of times over the course of the 3 three academic terms, station checklists for each examination are drawn from a large bank of available stations. Thus, the content of individual assessment checklists did not change over time.

3.4.2. Reliability Analysis

Table 3.2 illustrates two potential sources of variance allocated in the G-study accounting for more than 75% of the total variance. The MAAS-Global-by-checklists (nested within disciplines) interaction accounted for 39% of total error variance. Unidentified sources of variance (rest error) accounted for another 37% of total error variance. Moreover, MAAS-Global and disciplines-by-MAAS-Global interaction variance accounted for 8.6% and 9.1% of error variance respectively.

The results obtained from the G study demonstrated that the rater-by-checklists (nested within disciplines) and rater-by-disciplines interaction were the only source of error variance. The overall Generalized Kappa (G-Kappa) of the station checklists calibration, from all disciplines, was 0.8. D study analysis demonstrated a decrease in G-Kappa from 0.72 to 0.57 as the number of raters was reduced from 2 to 1.

3.4.3. G-Kappa Results for Discipline Checklists

To determine whether each station checklist contained CS items appropriate to the MAAS-Global checklist, a G-Theory analysis was carried out for each of the four disciplines [18]. G-Kappa results for the Disciplines of Obstetrics & Gynaecology, Paediatrics, General Practice, and Psychiatry were 0.45, 0.79, 0.80, and 0.99 respectively.

3.4.4. Communication Skills within Station Checklists

Table 3.3 presents a description of the content of each station checklist according to the MAAS-Global. The majority of the station checklists were related to section three of the MAAS-Global (46%) (Table 3.3, column 5). For sections one and two (column 3 and 4), the figures were 8% and 12% respectively. Meanwhile, 34% of checklist items were not considered to be CS items according to the MAAS-Global (column 6). The Discipline of Psychiatry was found to employ a higher percentage of items from section 3 of the MAAS-Global when compared to the other 3 disciplines.

It is apparent from this table that the Disciplines of Psychiatry and Obstetrics & Gynaecology did not assess items from different stages of the clinical consultation. Item 1 (Section 1) of the MAAS-Global refers to the initial phase of a clinical consultation, which focuses upon allowing the patient the opportunity to explain their presenting complaint. There were two items which were found to be used by all disciplines across all academic terms. These items are 'diagnosis' (Item 5, Section 1) and 'management strategy' (Item 17, Section 3). 'Information giving' (Item 10, Section 2) and 'measuring medical content of physical examination' (Item 15, Section 3) were found to be used by all disciplines across all academic terms, with the exception of the Discipline of Psychiatry in the 2011-2012 term. Table 3.3 also shows that the Disciplines of Paediatrics and General Practice incorporate the majority of the MAAS-Global items into their checklists. It is also noted that item 2 (i.e. follow-up consultation)

and item 7 (i.e. evaluation of consultation) of the MAAS-Global are not represented in any of the OSCE checklists.

3.5. Discussion and Conclusion

3.5.1. Discussion

This study set out to explore calibration of assessment forms used to assess CS, within and between disciplines in our School of Medicine. The MAAS-Global was used as a standard against which such assessment forms were compared. Considering the G-Kappa values, the study demonstrates that calibration of station checklists using the MAAS-Global as a standard is valid and reliable [23]. Validity in this respects refers to an evidence-based claim about the trustworthiness of decisions in CS assessment are based on the MAAS-Global standard and made from context-specific data [24]. In addition, the study demonstrates that use of this standard affords the opportunity to identify items which can be mapped to the MAAS-Global and this, subsequently, makes possible the comparison of CS assessments across different OSCE settings. This becomes possible as a result of the reliability and validity of the gold standard.

According to Van Es et al., the MAAS-Global is an instrument that assesses valid and reliable doctor–patient CS (patient-centered versus task-related skills) sustained in the dimensional structure of the MAAS-Global rating list for consultation skills in undergraduate and again postgraduate CS training [11,25]. In the present study, the MAAS-Global was used as the standard to calibrate our station checklists. It is apparent that those items which could be mapped to the MAAS-Global can be characterized as valid items for assessing relevant CS, whereas this may not be the case for items which could not be mapped to the standard.

Table 3.2: Summary of estimated variance component (G-Study), G-Kappa coefficient, and D Study (optimization) Analysis.

Source	df	Mean Squares	C o m p o n e n t	
Checklists (nested within Disciplines)	824	0.02186	0.00034	0.9 %
Disciplines	3	1.45460	0.00011	0.3 %
Raters	2	0.54890	0.00002	0.0 %
MAAS-Global	16	12.26489	0.00362	8.6 %
Rater * Checklists (nested within Disciplines)	1648	0.00430	0.00025	0.6 %
MAAS-Global * Checklists (nested within Disciplines)	13184	0.06113	0.01553	39.0 %
Disciplines * Raters	6	0.32270	0.00009	0.2 %
Disciplines * MAAS-Global	48	2.50588	0.00362	9.1 %
Raters * MAAS-Global	32	0.97495	0.00092	2.3 %
Raters * MAAS-Global * Disciplines	96	0.21141	0.00095	2.4 %
Raters * MAAS-Global * Checklists (nested within Disciplines)	26368	0.01455	0.01455	36.6 %

G-Kappa = 0.8
(Measurement design CD/RM) *

D Study (optimization) analysis:

Number of raters	G-Kappa
2	0.72
1	0.57

Generally, an acceptable reliability coefficient for high-stakes assessment is higher than 0.80, while 0.70 is acceptable for research purposes.

* C = checklists R = raters
D = Disciplines M = MAAS-Global

Table 3.3: MAAS-Global sections and items of stations' checklists in each discipline.

Disciplines	Academic Year	MAAS-Global Section (in percentage)				MAAS-Global Items (grouped in section)		
		1	2	3	N/A	1	2	3
Discipline of Obstetrics & Gynaecology	2010 - 2011	4	12	15	69	5	10, 13	15, 17
	2011 - 2012	8	11	12	69	5	10	15, 17
	2012 - 2013	3	8	36	53	5	10	14, 15, 17
Discipline of Paediatrics	2009 - 2010	15	16	50	19	1,3, 4,5,6,	8,10,13	14,15,17
	2010 - 2011	13	12	56	19	1,3, 4,5,6,	8,10,11,13	14,15,17
	2011 - 2012	9	7	47	37	1, 4, 5	8, 10	14, 15, 16, 17
	2012 - 2013	12	18	43	27	1, 4, 5, 6	8, 10, 11, 12, 13	14, 15, 16, 17
Discipline of General Practice	2009 - 2010	11	24	34	31	1, 3, 4, 6	8,10,11, 13	14,15,17
	2010 - 2011	12	22	41	25	1, 3, 4,5,6	8, 9,10,11,12,13	14,15,17
	2011 - 2012	10	23	40	27	1, 4, 5, 6	8, 9, 10, 11, 12, 13	14, 15, 16, 17
	2012 - 2013	14	23	39	24	1, 3, 4, 5, 6	8, 9, 10, 11, 12, 13	14, 15, 16, 17
Discipline of Psychiatry	2009 - 2010	6	12	76	6	5	10	14, 15, 16, 17
	2010 - 2011	5	6	81	8	5		14,15,17
	2011 - 2012	3	3	64	30	5		14,16, 17
	2012 - 2013	5	3	74	18	5	10	14, 15, 16, 17
Average (min-max) (all Disciplines)		8 (3-15%)	12 (3-24%)	46 (12-81%)	34 (8-69%)			

* N/A : not applicable

All raters were trained in the use of the MAAS-Global as a calibration tool. The term “raters-by- MAAS-Global interaction” is used in this G Study to describe the ability of raters to correctly use and interpret the MAAS-Global. The low level of error variance reported for the raters-by- MAAS-Global interaction demonstrates that raters have little difficulty in understanding the items described in the MAAS-Global when used as a calibration tool. The definitions of each of the MAAS-Global items, as outlined in the calibration manual, were well understood and applied to match station checklist items with MAAS-Global items. In this study, G Kappa was used to measure the level of agreement achieved when raters independently mapped each checklist item to MAAS-Global items. On moving from 3 raters to 2 raters, the G Kappa reduced by only 0.08, from 0.80 to 0.72, which implies that future calibration projects could be accurately carried out with only 2 independent raters.

The term “rater-by-checklist (nested within disciplines) interaction” is used in this G Study to describe the ability of raters to interpret checklists unique to each discipline OSCE station. The term “rater-by-discipline interaction” is used in this G Study to describe how the raters differ in their interpretation of checklists from different disciplines. We demonstrated that rater-by-disciplines interaction and rater-by-checklist (nested within disciplines) interaction were the only contributors to the calibration process error. These two sources of error show that raters are matching checklist items to MAAS-Global items differently when discipline specific CS items are involved. This may be due to variation in interpretation between raters from different professional backgrounds (i.e. a researcher, an educationalist and a clinician). The reader may assume that such difference would generate significant error, however the results of this study suggest that the single most important contributor to error is the way in which each discipline describes CS items in station checklists. Since station checklists were unique to disciplines, the level of discrepancy in agreement could be attributed to disciplines. This suggests that disciplines should exercise extreme

care in describing checklist items so that they are not misinterpreted by examiners or reviewers.

To support our findings, separate G Study analyses were conducted for each discipline. The result of G Kappa for the Discipline of Obstetrics & Gynaecology (0.45) was significantly below the conventionally accepted value of 0.8. This indicates that the raters had difficulty in matching checklist items with the MAAS-Global items. When the level of agreement between raters in their individual interpretation of checklist items was closely examined, it was apparent that raters differed significantly in their interpretation of what actually constituted a communication skill. In order to examine this phenomenon, all sections of the MAAS-Global were merged. Checklist items were then compared against this “merged MAAS-Global” in order to determine the level of agreement between raters in their identification of CS items. However, when re-calculated, the level of rater agreement (G Kappa) was essentially unchanged (0.44 vs 0.42). It is important to note that this result suggests significant variation in rater interpretation of checklist items, but does not however necessarily reflect the quality of the checklist.

The Discipline of Psychiatry had the highest G Kappa result (0.99). This result might be due to the fact that most checklist items in this discipline were easily categorized as section 3 items according to the MAAS-Global. Section 3 of the MAAS-Global addresses CS pertinent to medical history-taking, physical examination, diagnosis and management. One possible explanation for this result may be that it was relatively easy for raters to map each checklist item with this section of the MAAS-Global.

It was noteworthy that the Disciplines of Paediatrics and General Practice utilized the majority of MAAS-Global items; whilst the Disciplines of Obstetrics & Gynaecology and Psychiatry focused upon use of items from section 3 of the

MAAS-Global. This finding may result from sections 1 and 2 of the MAAS-Global having been assessed in earlier years of the programme. This finding merits further exploration and internal research.

The calibration procedure with 3 independent raters was labor intensive. The D-study shows the impact of lowering the numbers of raters' in future calibration procedures. Similar calibration procedures could be used within the consortium of users of our OSCE Management Information System [14]. The results show that calibrating assessment forms with only two raters is still a reliable process. Generalized Kappa for one or two raters is 0.57 and 0.72 respectively. Calibrating with only one rater is neither satisfactory nor realistic. Calibrating forms with two raters however is considered as an acceptable procedure e.g. acceptable reliability [18].

To be able to determine that CS education and assessment is occurring in a progressive fashion across the curriculum, further studies need to be undertaken using checklists from all stages of the programme of study. We have carried out a vertical cross-section analysis of OSCE CS items used in the assessment of consecutive cohorts of year 4 students. Suggested future research should include a horizontal comparison across the entire programme of study so that progressive change in CS outcomes can be identified. It is assumed that it is possible to assess different sections of the MAAS-Global, such as CS for each separate phase of consultation, general CS, or the medical aspect of CS across different years (i.e. different levels/stages of CS). It is apparent that in year 4, emphasis is on the medical content during medical consultations (i.e. history taking, physical examination, diagnosis, and management). It is not known whether other phases of the consultation and other generic CS are being appropriately assessed in earlier years of the degree programme (i.e. years 1, 2 and 3). A further suggestion for future research is to explore the possible

measurement of 'change in CS' over time, using the smallest detectable difference (SDD) [19].

Rather than repeatedly creating new assessment forms, researchers and educators should work together in order to agree upon the definitions of learning outcomes and CS domains to be assessed. A clear description of the learning objectives, or underlying concepts of assessment forms being used, is frequently absent. Items of CS to be assessed need to be mapped to an existing universally accepted standard for CS and they also need to be mapped to learning outcomes [26]. Great emphasis on this set of skills in relation to the attainment of professional competencies is laid out by regulatory bodies worldwide [17]. It is our professional duty to ensure that our assessments, and their results, are defensible and that our assessment forms are sensitive enough to discriminate between 'good' and 'bad' performance and to measure change over time [19].

3.5.2. Limitation of the Study

The calibration of 2577 items proved to be extremely labour intensive. Calibration in terms of mapping OSCE station items with either a standard or CS training learning outcomes should be conducted by content experts prior to the design of new OSCE forms rather than after the OSCE has taken place. The OSCE Management Information System could be adjusted in such a way that mapping with curriculum outcome measures would be possible. In that case, not only would it be possible to produce an instant analysis of the outcomes of CS training, but it would also be possible to map these against any available standard or competency model. It is also acknowledged that the present study could not take in to consideration any change in learning activities that may have taken place over the course of the study period. Regarding the internal validity of the CS checklists reviewed in the present study, we acknowledge that it is important to also be aware of the purpose of each individual CS station in order

to enable determination of construct validity and relevance of each OSCE station checklist. The raters in the present study did not have access to this additional information.

3.5.3. Conclusion

In the present study, station checklist items were calibrated and categorized according to the MAAS-Global. Significant heterogeneity in approach to the assessment of CS was identified between different disciplines. The calibration of OSCE checklist items, according to the MAAS-Global, is possible and the procedure was been shown to be reliable. This study thereby provides supportive evidence for using the MAAS-Global checklist as a tool to calibrate different types of CS items in OSCE station checklists. Such calibration will enable comparison of results of CS assessments between students and across different discipline-specific learning outcomes. By transforming OSCE checklist scores into grades that are standardized against the MAAS-global, standardized comparison between and within cohorts of students becomes feasible and will be the subject of our future research. We suggest that the MAAS-Global be more widely employed as a calibration tool. Future research should focus upon exploration of the progress of CS assessment and CS outcomes across an entire programme of study.

3.6. Practice Implications

It is possible to compare OSCE checklist items against an agreed gold standard and thereby measure the 'true' caliber of CS in OSCE stations. In that way, results can be compared between and across stations, students and institutions. We suggest that future OSCE station design should be more carefully blueprinted against the curriculum so that assessments match with CS learning outcomes. With regards to generalizability of results, reliable calibration procedures require only two instead of three raters (G -coefficient = 0.72). Our quality assurance process employs both instant outcome analysis of OSCE assessments and the

implications of this research to improve station design. We suggest that an alternative approach would be the de novo design of “MAAS-Global OSCE CS stations”, which directly assess items from the MAAS-Global.

3.7. References

- [1] Von Fragstein M, Silverman J, Cushing A, Quilligan S, Salisbury H, Wiskin C, et al. UK consensus statement on the content of communication curricula in undergraduate medical education. *Med Educ* 2008;42:1100–7. doi:10.1111/j.1365-2923.2008.03137.x.
- [2] Boon H, Stewart M. Patient-physician communication assessment instruments:: 1986 to 1996 in review. *Patient Educ Couns* 1998;35:161–76. doi:10.1016/S0738-3991(98)00063-9.
- [3] Beck RS, Daughtridge R, Sloane PD. Physician-patient communication in the primary care office: a systematic review. *J Am Board Fam Pract* 2002;15:25–38.
- [4] Cegala DJ, Lenzmeier Broz S. Physician communication skills training: a review of theoretical backgrounds, objectives and skills. *Med Educ* 2002;36:1004–16.
- [5] Setyonugroho W, Kennedy KM, Kropmans TJB. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Educ Couns* (2015), <http://dx.doi.org/10.1016/j.pec.2015.06.004>.
- [6] Guiton G, Hodgson CS, Delandshere G, Wilkerson L. Communication skills in standardized-patient assessment of final-year medical students: a psychometric study. *Adv Health Sci Educ* 2004;9:179–87.
- [7] Kiessling C, Essers G, Anvik T, Jankowska K. General principles for the assessment of communication skills 2015. (http://www.each.eu/wp-content/uploads/2014/07/General_principles_for_the_assessment_of_communication_skills_final.pdf): (accessed 4.26.2015).
- [8] Carmines EG, Zeller RA. *Reliability and Validity Assessment*. SAGE; 1979.
- [9] Schirmer JM, Mauksch L, Lang F, Marvel MK, Zoppi K, Epstein RM, et al. Assessing communication competence: a review of current tools. *Fam Med* 2005;37:184–92.
- [10] Roter DL, Hall JA, Blanch-Hartigan D, Larson S, Frankel RM. Slicing it thin: new methods for brief sampling analysis using RIAS-coded medical dialogue. *Patient Educ Couns* 2011;82:410–9. doi:10.1016/j.pec.2010.11.019.
- [11] Van Dalen J, Kerkhofs E, van Knippenberg-van den Berg BW, van den Hout HA, Scherpbier AJJA, van der Vleuten CPM. Longitudinal and concentrated communication skills programmes: Two Dutch medical schools compared 2002.
- [12] Silverman J. The Calgary-Cambridge guides: the “teenage years.” *Clin Teach* 2007;4:87–93.

- [13] J van D. d. *Educ Health* 2007;20:88. (<http://www.educationforhealth.net/article.asp?issn=1357-6283;year=2007;volume=20;issue=2;spage=88;epage=88;aulast=van;type=0>): (accessed 4.25.2015).
- [14] Kropmans TJ, O'Donovan BG, Cunningham D, Murphy AW, Flaherty G, Nestel D, et al. An Online Management Information System for Objective Structured Clinical Examinations. *Comput Inf Sci* 2011;5:p38. doi:10.5539/cis.v5n1p38.
- [15] Khan KZ, Gaunt K, Ramachandran S, Pushkar P. The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration. *Med Teach* 2013;35:e1447–63. doi:10.3109/0142159X.2013.818635.
- [16] Van Thiel J, Ram P, van Dalen J. MAAS-global manual. Maastricht Maastricht Univ 2000:4–5. (http://www.hag.unimaas.nl/Maas-global_2000/GB/MAAS-Global-2000-EN.pdf): (accessed 8.11.2012).
- [17] Brown J. How clinical communication has become a core part of medical education in the UK. *Med Educ* 2008;42:271–8. doi:10.1111/j.1365-2923.2007.02955.x.
- [18] Cardinet J, Johnson S, Pini G. *Applying Generalizability Theory Using Edug*. Taylor & Francis; 2012.
- [19] Kropmans T, Dijkstra P, Stegenga B, Stewart R, De Bont L. Smallest detectable difference of maximal mouth opening in patients with painfully restricted temporomandibular joint function. *Eur J Oral Sci* 2000;108:9–13.
- [20] Boyko JA, Lavis JN, Dobbins M, Souza NM. Reliability of a tool for measuring theory of planned behaviour constructs for use in evaluating research use in policymaking. *Health Res Policy Syst* 2011;9:29. doi:10.1186/1478-4505-9-29.
- [21] Lakes KD, Hoyt WT. Applications of Generalizability Theory to Clinical Child and Adolescent Psychology Research*. *J Clin Child Adolesc Psychol* 2009;38:144–65. doi:10.1080/15374410802575461.
- [22] Briesch AM, Swaminathan H, Welsh M, Chafouleas SM. Generalizability theory: A practical guide to study design, implementation, and interpretation. *J Sch Psychol* 2014;52:13–35. doi:10.1016/j.jsp.2013.11.008.
- [23] Bloch R, Norman G. Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68. *Med Teach* 2012;34:960–92. doi:10.3109/0142159X.2012.703791.
- [24] Andreatta PB, Marzano DA, Curran DS. Validity: what does it mean for competency-based assessment in obstetrics and gynecology? *Am J Obstet Gynecol* 2011;204:384.e1–384.e6. doi:10.1016/j.ajog.2011.01.061.
- [25] Van Es JM, Schrijver CJW, Oberink RHH, Visser MRM. Two-dimensional structure of the MAAS-Global rating list for consultation skills of doctors. *Med Teach* 2012. doi:10.3109/0142159X.2012.709652.
- [26] Von Fragstein M, Silverman J, Cushing A, Quilligan S, Salisbury H, Wiskin

C, et al. UK consensus statement on the content of communication curricula in undergraduate medical education. *Med Educ* 2008;42:1100–7.
doi:10.1111/j.1365-2923.2008.03137.x.

Chapter 4. True Communication Skills Assessment in Interdepartmental OSCE Stations : Comparison of Multiple Instruments

This chapter was submitted for publication in Education for Health, 2015.

4.1. Abstract

Background. Comparing uncalibrated outcome of clinical skills assessment is challenging. This study proposes reliable and valid comparison of communication skills (CS) assessment as practiced in two OSCEs originating from different disciplines. Our aim is to determine the differences in CS assessment between year 4 OSCE stations in an undergraduate Irish medical school in order to identify the characteristics of each test, using calibrated assessment forms.

Method. Three academic terms of the year 4 CS OSCE were analysed. The OSCE employed interdiscipline Psychiatry and General Practice stations. We chose the MAAS-Global as an internationally recognized and validated instrument to calibrate the OSCE station items. The MAAS-Global proportion is the percentage of station checklist items that can be considered as 'true' CS. Initial station scores were multiplied by the MAAS-Global proportion in order to obtain MAAS-Global scores. The reliability of the OSCE was calculated with G-Theory analysis and nested ANOVA was used to compare mean scores of all years.

Results. The calibrated outcome of Psychiatry stations demonstrated higher Maas-Global scores than General Practice stations. The proportion of section 3 items of the MAAS-Global (medical content) was larger than the proportion section 1 (sequential) and 2 (generic) items. We used average (sd) raw OSCE scores of six circuits over the past three years. General Practice station scores ranged from 58 (sd=6) to 64 (sd= 6). MAAS -Global scores ranged from 44 (sd=4) to 48 (sd=5). For Psychiatry, station scores ranged from 61 (sd=8) to 70 (sd=10) whereas MAAS-Global scores ranged from 59 (sd=8) to 66 (sd=8). MAAS-Global

scores in Psychiatry stations were significantly higher ($p < 0.03$) and above the initial passmark of 50%.

Conclusion. The higher students' scores in Psychiatry stations was a direct result of higher MAAS-Global proportion compared to the General Practice stations. Comparison of outcome measurements between interdiscipline station checklists was valid and reliable. The MAAS-Global was used as a single validated instrument and is suggested as gold standard. Future research needs to determine what to do about CS scores in 'mixed stations' that appears to be below the pass mark of 50%.

4.2. Background

Comparing assessment outcome of communication skills (CS) in undergraduate medical education is challenging as it is intended to measure differences between individual students, but is also influenced by cohorts of students and differences in learning outcomes. A synthesis of the literature demonstrates that a large proportion of medical errors and adverse events in postgraduate professional practice can be explained by communication factors between clinicians and patients and between health care providers themselves [1,2]. According to Elder and Dovey, 2002, other errors are categorized as administration factors (office and personnel issues), and blunt end factors [3]. Schoenthaler et al mentioned in her recent systematic review that only few papers targeted patient–practitioner communication and assessed the impact on cardiovascular-related clinical outcomes, limiting the ability to determine effectiveness in other areas such as well-being and compliance [4]. Referring to previous research, we support Schoenthalers' conclusion that additional rigorous research supported by theoretical frameworks and validated measurement is required to understand the potential of patient–practitioner communication to improve cardiovascular-related clinical outcomes. Communication appears also to be extremely important to ensure safe and effective clinical practice in

operation theatres. Communication was shown to influence operating theatre practices in all studies [5]. Further detailed observational research is needed to gain a better understanding of how to improve the doctors' working environment and patient safety. Training and assessment of clinical communication starts at undergraduate level and the purpose of this training is to adequately prepare undergraduate students for professional practice. This appears to be a major challenge for educational institutions. Nevertheless, with a careful prospective design, an OSCE may be developed to simultaneously assess multiple competencies including CS. Holistic ratings as well as checklists may help to evaluate physician competencies (CanMeds) in a reliable and valid manner in the OSCE.

In previously published research we explored 27 domains of CS being assessed and discussed in the international medical education literature [6]. Comparing the existing evidence about reliable and valid assessment to our existing assessment practice, suggests to us that assessment of CS with either a checklist or global rating scale is like measuring these crucial skills using a rubber band because those measurement instruments are not standardized. We previously attempted to calibrate all our existing clinical skills assessment forms used in previous OSCEs to assess 'aspects of CS' and adopted the MAAS-Global as a validated and reliable assessment tool and the "gold standard" within our Medical School [7,8]. In previous research, we calibrated all clinical skills items of our OSCE forms in the year 4 OSCEs to arrive at adequate comparison of the CS component of each OSCE station [7]. According to Van Es et al., the MAAS-Global has been proven to be a valid and reliable instrument [9].

In the present study, we introduce a conversion method using the MAAS-Global percentage score of our initial raw scores to assess the true CS measured with our existing station checklists. It is apparent that those items that could be mapped to the MAAS-Global can be characterized as valid items to assess 'true'

CS, whereas this may not be the case for items which could not be mapped to a standard. This study proposes reliable and valid comparison of CS by comparing 'raw scores of communication skills' measured with 'uncalibrated' assessment forms with calibrated 'true' measurement of CS [7].

The aim is to determine the differences in CS assessed in a penultimate year OSCE stations in an undergraduate medical school and identify the characteristic of each test result using assessment forms that were calibrated with the MAAS-Global.

4.3. Methods

4.3.1. Context of the Study

This retrospective study analyzed the penultimate OSCE of the undergraduate medical programme at the National University of Ireland in Galway, Ireland. The data from the station forms developed by the Discipline of Psychiatry and those of the Discipline of General Practice were retrieved from our online OSCE Management Information System (OMIS). CS are commonly measured in an OSCE where students' ability to communicate with the standardized patient is measured by examiners. Unlike language tests which already are measured with standardized tests, e.g. TOEFL, CS do not have such a standardized test. Direct comparison of CS attained from station results is not possible, since every station in general uses different station checklists.

4.3.2. Research Subjects

This study analyzed data from three academic terms, 2010/2011, 2011/2012, and 2012/2013. The penultimate OSCE is administered in February/March and April, for two consecutive groups of students. Total students for academic terms 2010/2011, 2011/2012, 2012/2013 were 116, 123, and 141 respectively (total n=380).

Both disciplines use different OSCE settings, e.g. number of stations, sequence of stations, or scoring rubrics. The station durations were 5 minutes for General Practice stations and 10 minutes for Psychiatry stations.

4.3.3. Measurement Instrument

The MAAS-Global is designed as a generic instrument to rate physicians' CS. The MAAS-Global consists of 17 items divided into 3 sections. Seven items in section 1 refer to appropriate skills in the specific phases of clinical consultations. Items are related to introduction, follow-up consultation, a request for help, physical examination, diagnosis, management, and evaluation of the consultation. These items are a reflection of the logical order of consultation phases. Section 2 focuses on general CS which occur throughout the consultation, consisting of 6 items. Those items are: exploration, emotions, information giving, summarizations, structuring, and empathy. Section 3 is intended to examine the mastery of the medical content during medical consultation. This section consists of 4 items: history taking, physical examination, diagnosis, and management which represent phases of the consultation.

4.3.4. Conversion Method

The conversion method used in this study is a process where the unstandardized measurement instruments are calibrated with the MAAS-Global, hence allowing the examination results to become comparable between students, disciplines and different groups.

The first step of this method is to calibrate each measurement instrument i.e. station checklist, with the MAAS-Global. The calibration method has been proven to be a reliable procedure according to our previous study (Setyonugroho 2, 2015). This calibration result is what we will call the MAAS-Global proportion, which is a percentage amount of checklist items that are considered to be CS items according to the MAAS-Global. The second step of this conversion method

is converting the students' score. Students' scores in each station multiplied by MAAS-Global proportion is the MAAS-Global score. The total MAAS-Global score of the OSCE is the average of the station's MAAS-Global score over all relevant OSCE stations assessing any type of CS.

In conditions where not all of the OSCE stations have CS items, the total MAAS-Global score is the average of station with MAAS-Global score only (i.e. in April 2012, from 5 stations in Discipline of Psychiatry, only of 4 stations a MAAS-Global score could be calculated. Hence the total MAAS-Global score is average of 4 stations' MAAS-Global score).

4.3.5. Statistical Analysis

The results of each OSCE were combined and converted to a percentage scale. Generalizability Theory was calculated independently for each circuit of each discipline to determine reliability of the test. Nested ANOVA was used for the comparison of the mean of six circuits and Tukey's post-hoc test was used to distinguish which of the circuits introduced difference. Statistical significance was set at the .03 level of probability. A software package EduG (Swiss Society for Research in Education Working Group. Edumetrics - Quality of measurement in education) has been used to perform G-Theory analysis and SPSS version 21 was used to analyze ANOVA.

4.4. Results

4.4.1. Station Characteristics

The settings of OSCE circuits in both General Practice and Psychiatry parts of the OSCE are different. For General Practice stations, the number of stations in all circuits in all academic terms is 10. The station duration is 5 minutes for each station. Meanwhile, for Psychiatry stations, each circuit in term 2010/2011 incorporates 4 stations while in terms 2011/2012 and 2012/2013 each circuit

consists of 5 stations. The station duration of psychiatry stations is twice as long as those in General Practice (i.e. 10 minutes each).

As was pointed out previously, we calibrated all station checklists in both disciplines using MAAS-Global as a standard. The calibration revealed that 9 out of 10 General Practice stations contain CS, except for one held in February of the 2010 – 2011 terms. In Psychiatry, all stations in 2010 - 2011 contain CS while one station did not contain CS in both 2011 - 2012 and 2012 – 2013 terms.

Table 4.1 presents the characteristics of the station's checklists of General Practice and Psychiatry stations. In the General Practice stations, the MAAS-Global proportion range between 65 to 75 percent, with an average of 68 for 3 academic terms. It is apparent from Table 1 that section 3 of the MAAS-Global (mastery of the medical content during consultation), is the largest portion of the CS assessed according to MAAS-Global. In section 2 (generic skills) between 29 to 35 percent and of section 1 (sequential skills) between 11 to 19 percent were 'true' CS items.

Table 4.1 also demonstrates that the General Practice' stations incorporate almost all of the MAAS-Global items in all circuits. By contrast, the Discipline of Psychiatry incorporate only item 5 (diagnosis) and item 10 (information giving) of section 1 and 2. Moreover, only 10 percent or less of the Psychiatry stations was not considered as communication skills (2010/2011), which was shown by high MAAS-Global proportion (90 and 92 percent). While in both 2011/2012 and 2012/13 terms, the MAAS-Global proportion is 70 percent or more. It is apparent of this dataset that the majority of the station checklist items are considered as section 3 of MAAS-Global items (range 77 to 100 percent).

Table 4.1: Summary of OSCE score Mean, MAAS-Global score Mean, Communication skills, MAAS-Global proportion, section of MAAS-Global in percentage, and MAAS-Global items.

Discipline of General Practice										
Academic Year	OSCE Circuit	OSCE Mean (SD)*	MG Mean (SD)*	MG Proportion	MAAS-Global (%)			MAAS-Global Items		
					Section 1	Section 2	Section 3	Section 1	Section 2	Section 3
2010/2011	February	58 (5.5)	43.8 (4)	75	14	29	57	1,3,4,5,6	8,9,10,11,12,13	14,15,17
	April	58.6 (4.9)	45.8 (3.6)	69	19	29	52	1,4,5,6	8,9,10,11,12,13	14,15,17
2011/2012	February	63.2 (4.3)	43.9 (3.3)	65	16	35	49	1,4,5,6	8,9,10,11,12,13	14,15,17
	April	61.3 (5.8)	46.6 (4.2)	67	11	35	54	1,4,5,6	8,9,10,11,12,13	14,15,17
2012/2013	February	63.9 (6.3)	48.2 (4.5)	66	14	27	59	1,4,5,6	8,9,10,11,12,13	14,15,17
	April	61.6 (5.8)	47.1 (4.4)	71	15	30	55	1,3,4,5,6	8,9,10,11,12,13	14,15,17
Discipline of Psychiatry										
2010/2011	February	69.8 (9.9)	62.6 (8.9)	90	13	6	81	5	10	14,17
	April	64.6 (8.8)	60.4 (8.2)	92	0	0	100			14,15,17
2011/2012	February	68.9 (8.3)	66.2 (8.2)	73	2	4	94	5	10	14,15,16,17
	April	61.2 (8.4)	60.6 (7.5)	74	5	6	89	5	10	14,16,17
2012/2013	March	69.1 (9.9)	61.4 (8.6)	70	16	7	77	5	10	14,17
	April	63.1 (8.5)	58.5 (7.9)	73	1	4	95	5	10	14,16,17

* score scale 1-100
 abbreviation :
 MG : MAAS-Global
 SD : standard deviation

4.4.2. Scores

Originally, the MAAS-Global score is a rating of CS. In this study, the result of a calibrated OSCE using the MAAS-Global is also called MAAS-Global score. Therefore, from this point onward we will use the terms MAAS-Global score followed by MAAS-Global proportion and section percentage. For example, MAAS-Global score 65 with MAAS-Global proportion 75 which consist of 14% section 1, 29% section 2 , and 57% section 3, will be then written as MAAS-Global score 65 [MG75-14-29-57].

From the data in Table 4.1, it can be seen that average OSCE scores (scale 0-100) of the six circuits of the Discipline of General Practice ranged from 58 (sd=5.5) to 63.9 (sd= 6.3). The MAAS-Global scores average ranged from 43.8 (sd=4) [MG75-14-29-57] to 48.2 (sd=4.5)[MG66-15-30-55].

Whereas the average OSCE scores of the Discipline of Psychiatry ranged from 61.2 (sd=8.4) to 69.8 (sd=9.9). The MAAS-Global scores average of the Psychiatry stations tended to be higher, ranging from 58.5 (sd=7.9) [MG73-1-4-95] to 66.2 (sd=8.2)[MG73-2-4-96].

A Nested ANOVA was conducted to compare the effect of disciplines and circuits which nested within disciplines. Table 4.2 shows the analysis of variance for both OSCE score and MAAS-Global score.

For the OSCE score, there was a significant effect for discipline at the $p < .03$ level [$F(1, 748) = 84.12, p < .001$] and there was a significant effect for circuit which was nested within the discipline $F(10, 748) = 10.58, p < .001$. While for MAAS-Global scores we found similar results, there was significant effects for discipline and circuit which was nested within the discipline, $F(1, 748) = 1080.83, p < .001$ and $F(10, 748) = 7.50, p < .001$ respectively. So results of CS measurements differ statistically significant between discipline stations and measured in various circuits.

Table 4.2: Analysis of variance for both OSCE score and MAAS-Global score (compare the effect of disciplines and circuits which nested within disciplines).

OSCE score			
Effect	Sum of Squares	df	Mean Square
Discipline	4694.63	1	4694.63
Circuit nested within Discipline	5909.67	10	590.97
Error	41745.01	748	55.81
MAAS-Global score			
Effect	Sum of Squares	df	Mean Square
Discipline	45342.89	1	45342.89
Circuit nested within Discipline	3148.7	10	314.87
Error	31380.04	748	41.95

To determine which circuits caused the difference, one-way ANOVA was calculated independently for each discipline to compare the effect of different circuits (six circuits in three academic years). The means and standard deviations are presented in Table 4.1.

For the Discipline of General Practice, at the $p < .03$ level, there were significant differences between the circuits in OSCE scores [$F(5, 379) = 11.16, p < .001$] and MAAS-Global scores [$F(5, 379) = 11.99, p < .001$]. The Tukey comparison in OSCE scores revealed that results achieved over circuits being assessed in 2010/2011 were statistically significantly lower than results achieved in circuits for both terms 2011/12 and 2012/2013. While for MAAS-Global scores, the Tukey comparison shows that for the February 2010/2011 OSCE, the April 2010/2011 OSCE, and the February 2011/2012 OSCE were significantly lower than the other three circuits in those years.

In Discipline of Psychiatry, at the $p < .03$ level, OSCE scores were significantly different for different circuits [$F(5, 379) = 10.37, p < .001$] and MAAS-Global

scores [$F(5, 379) = 6.41, p < .001$] (see Table 4.1 for means). The Tukey comparison in OSCE scores revealed that circuits in February/March were significantly higher than April. While in MAAS-Global scores, April 2010/2012 and 2012/2013 were significantly higher than 4 other circuits.

4.4.3. Reliability Analysis

A generalizability analysis was performed separately for each circuit of both disciplines. From data in Table 4.2, it can be seen that in 2011/2012, the February circuits run by the Discipline of General Practice has the lowest G-coefficient .59, while in other circuits the generalizability is just over 0.7.

In a Decision Study (D-study), a hypothetical design with 10 stations of Discipline Psychiatry OSCE was calculated. The results for the year 2010/2011 to 2012/2013 (February and April) would have been 0.74, 0.79, 0.79, 0.72, 0.84, and 0.81, respectively, if 10 stations were used. Table 4.3 shows a study design with 15 stations for both disciplines. In that case the majority of the G-coefficients in both disciplines majority would have been 0.80 or more.

Table 4.3: Summary of Generalizability Coefficient and Decision Study (with 10 and 15 stations).

Academic Terms	Circuit	Discipline of General Practice		Discipline of Psychiatry		
		G	D-study (15)*	G	D-study (10)*	D-study (15)*
2010/2011	February	0.75	0.82	0.54	0.74	0.85
	April	0.73	0.8	0.61	0.79	0.85
2011/2012	February	0.59	0.68	0.65	0.79	0.85
	April	0.73	0.8	0.56	0.72	0.79
2012/2013	February	0.74	0.81	0.73	0.84	0.89
	April	0.7	0.77	0.68	0.81	0.87

abbreviation list :

D-Study : Decision Study

G : G coefficient

4.5. Discussion

The aim of the study was to determine the 'true' differences in CS outcomes as assessed in a penultimate OSCE of undergraduate medical students comparing 'raw' scores and 'MAAS-Global' scores. We identified the characteristics of each test result using assessment forms that were calibrated with the MAAS-Global. The reliability according to the G-theory was moderate to good and could be improved by increasing the number of stations to at least 10 for the Psychiatry OSCE. It is apparent that 10 stations, each of 5 minutes duration, generate a higher G-coefficient than 5 stations of 10 minutes duration. This supports the concept mooted in previous publications that a larger number of stations lead to higher reliability [10]. Our D-study for 15 stations for both disciplines revealed that the reliability of Discipline of Psychiatry is relatively higher than Discipline of General Practice. The possible explanation for higher reliability in Discipline of Psychiatry may be that the station duration is longer than those for the Discipline of General Practice [11].

There is an apparent difference in MAAS-Global items being covered between stations developed by General Practice and those developed by the Discipline of Psychiatry. Whilst in General Practice learning outcomes of CS training are more generic and apparently cover different stages of the consultation (Section 1 items of the MAAS-Global) being a reflection of the logical order of consultation phases with items related to introduction, follow-up consultation, a request for help, physical examination, diagnosis, management, and evaluation of the consultation. Section 2 focuses on general CS which are used throughout the consultation, consisting of 6 items. Those items are: exploration, emotions, information giving, summarizations, structuring, and empathy. In the Discipline of Psychiatry stations, there was an emphasis upon the medical content of the consultation, in contrast with the Discipline of General Practice, where there was a broader emphasis upon a wider range of CS items as defined by the MAAS-

Global. In the Discipline of Psychiatry, there were more CS items that were specifically addressing history taking, physical examination, diagnosis, and management of diseases. As it is difficult to map the learning outcomes of CS training for disciplines' 'CS assessment forms' it is not certain whether the average coverage of 15, 31 and 55% MAAS-Global sections respectively covers learning outcomes related to CS that specifically focus on different stages of the consultation (section 1), general CS (section 2) and mastering the medical content of the consultation (section 3). Although we have automated clinical skills assessments in our School of Medicine, and we are able to calibrate our assessment forms in an automated fashion, we do not yet have a curriculum mapping tool linking specific items or competencies with the curriculum learning outcomes. It is obvious that Psychiatry stations with respectively 6, 5 and 89% MAAS-Global sections are assessing something different to General Practice stations. Furthermore, General Practice was found to be assessing a combination of communication (69%) and technical skills (31%), whereas Psychiatry is predominantly assessing CS (79%). In previous research the Generalized Kappa for reviewers agreement about calibration of forms was high for the General Practice stations (0.83), but was even higher for the Psychiatry stations (0.99) [7]. Thus, we can be quite sure that the Psychiatry stations are assessing section 3 CS items, whereas the nature of the General Practice CS items is more open for debate.

Regarding content validity, a typical General Practice consultation entails both technical skills and CS, however in a learning situation it is necessary to confirm whether our assessment strategy verifies that students have achieved their learning goals and whether standard setting has been successful [12]. The pass mark in Schools of Medicine in Ireland is generally regarded as 50%. Whilst pass marks vary between universities worldwide, there is not much evidence available to confirm the validity or ratio behind the use of static pass marks. With the OSCE Management Information System, we use Borderline Regression Analysis

to incorporate the difficulty of stations and variability between examiners. Marks are presented in terms of a regression outcome and, also, as a static pass mark of 50% [13]. Nevertheless, the average (SD) MAAS-Global score for General Practice stations is 46 (4) versus 62 (8) for Psychiatry measured over three academic terms. These scores indicate that 68% of our cohort of students being assessed achieved a 'true' score for CS between 42 (average minus 1 standard deviation) and 50% (average plus 1 standard deviation) which would indicate a performance below formal standard setting. For the Psychiatry stations, in which 68% of students of different cohort(s) achieved a score between 54 and 70, this is less significant an issue than in General Practice. A MAAS-Global score below 50% does not indicate a fail score. The MAAS-Global score was attained from OSCE score multiplied with MAAS-Global proportion, hence the actual maximum MAAS-Global score is similar to MAAS-Global proportion i.e. not equal to 100%. Furthermore, as explained in Table 4.1, each OSCE had a different MAAS-Global proportion. International literature pertaining to OSCE assessments addresses tendencies toward competency based marking whereas the assessment forms analyzed in the present study are very much item based 'tick box' assessment forms. Further research should consider matching the various intersections of our assessment forms addressing respectively section 1, 2 and 3 items of the MAAS-Global in order to generate a 'competency based score' for CS [7]. With respect to this dataset, we suggest the use of an overall MAAS-Global score as outlined in Table 4.1, 2nd row, column 4 – 8 being 43.8 (4) MAAS Global proportion 75%; with 14 % section 1; 29% section 2 and 57% section 3 items i.e. MAAS-Global score 44 [MG75-14-29-57], indicates a score below the pass mark of 50% in an OSCE addressing 75% CS and 25% technical (other) skills, not being CS.

“Fixing the rubber band” to measure CS is not only about calibrating assessment forms, it is also about the requirement of assessment forms to be able to measure change. Sensitivity to change in clinimetrics is a well-established area of

research. Sensitivity to change is part of the classical psychometric analysis of diagnostic, prognostic and therapeutic discriminative and evaluative instruments. There is however a notable shortage of assessment tools and psychometric characteristics to be used for progress monitoring in clinical skills assessments. Considerable evidence is available to measure growth in medical knowledge [13]. Recently, Turan and Valcke c.s. (2013) developed a Medical Achievement Self-efficacy Scale (MASS) for students of the Ghent Curriculum [15]. The latter scale is related to the general competency frameworks of CanMEDs and the Five-star Doctor and predict progress test outcome, however clinical skills assessments are not included. Multiple mini-interviews predict clerkship and licensing examination performance, including OSCE performance, but again evidence regarding the measurement of change in clinical skills assessments is lacking [16]. Feedback is believed to a most powerful learning tool. However, we should be able to measure change and being able to measure 'statistical significant change in the individual subject' to verify whether our teaching and learning intervention has been successful in individual students.

The procedure of calculating MAAS-Global scores based on the MAAS-Global standard is labour-intensive. This process currently entails two steps, calibrating the OSCE station-checklists and re-calculating each students MG score for each of the stations. To date our OSCE Management Information System does not have an automated mechanism for this process. Future development of the software could incorporate the option to map OSCE rubrics and calculate the MG score directly. However, it might be easier to develop specific MAAS-Global based CS stations. Moreover, future research should consider possible use of the MG score as one of the criteria for standard setting of an OSCE. In such a case, not only would students have to pass the overall OSCE cut-score, but they would also have to pass a minimum MG score.

4.6. Conclusion

This study has demonstrated the process for distilling an MG score from an overall OSCE score. Secondly, we demonstrated the true characteristics of CS based on a standardized instrument (i.e. the MAAS-Global). Returning to the research question posed at the beginning of this study, it is now possible to compare CS assessment outcomes from different settings (i.e. rubrics or different modules) of OSCEs. Moreover, this new approach should be considered as a possible standard procedure to assess CS in OSCEs and to improve quality of measurement. Future research should be undertaken to explore how to incorporate the 'true' CS score as one criterion for passing the conjunctive standard.

4.7. References

- [1] Phillips C. Communication: The First Tool in Risk Management for Long-Term Care. *J Am Med Dir Assoc* 2004;5:123–6. doi:10.1016/S1525-8610(04)70067-5.
- [2] Beck RS, Daughtridge R, Sloane PD. Physician-patient communication in the primary care office: a systematic review. *J Am Board Fam Pract* 2002;15:25–38.
- [3] Elder NC, Dovey SM. Classification of medical errors and preventable adverse events in primary care: a synthesis of the literature. *J Fam Pract* 2002;51:927–32.
- [4] Schoenthaler A, Kalet A, Nicholson J, Lipkin M. Does improving patient-practitioner communication improve clinical outcomes in patients with cardiovascular diseases? A systematic review of the evidence. *Patient Educ Couns* 2014;96:3–12. doi:10.1016/j.pec.2014.04.006.
- [5] Weldon S-M, Korkiakangas T, Bezemer J, Kneebone R. Communication in the operating theatre: Communication in the operating theatre. *Br J Surg* 2013;100:1677–88. doi:10.1002/bjs.9332.
- [6] Setyonugroho W, Kennedy KM, Kropmans TJB. Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. *Patient Educ Couns* 2015. doi:10.1016/j.pec.2015.06.004.
- [7] Setyonugroho W, Kropmans T, Kennedy KM, Stewart B, Dalen J van. Calibration of Communication Skills Items in OSCE Checklists according to the MAAS-Global. *Patient Educ Couns* 2015. doi:10.1016/j.pec.2015.08.001.
- [8] Van Es JM, Schrijver CJW, Oberink RHH, Visser MRM. Two-dimensional

structure of the MAAS-Global rating list for consultation skills of doctors. *Med Teach* 2012. doi:10.3109/0142159X.2012.709652.

[9] Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. *Med Educ* 2011;45:1181–9. doi:10.1111/j.1365-2923.2011.04075.x.

[10] Smith V, Muldoon K, Biesty L. The Objective Structured Clinical Examination (OSCE) as a strategy for assessing clinical competence in midwifery education in Ireland: A critical review. *Nurse Educ Pract* 2012;12:242–7. doi:10.1016/j.nepr.2012.04.012.

[11] Ben-David MF. AMEE Guide No. 18: Standard setting in student assessment. *Med Teach* 2000;22:120–30. doi:10.1080/01421590078526.

[12] Hejri SM, Jalili M, Muijtjens AMM, Van Der Vleuten CPM. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *J Res Med Sci Off J Isfahan Univ Med Sci* 2013;18:887–91.

[13] Muijtjens AMM, Schuwirth LWT, Cohen-Schotanus J, Thoben AJNM, Vleuten CPM van der. Benchmarking by cross-institutional comparison of student achievement in a progress test. *Med Educ* 2008;42:82–8. doi:10.1111/j.1365-2923.2007.02896.x.

[14] Turan S, Valcke M, De Maeseneer J, Aper L, Koole S, De Wispelaere C, et al. A novel Medical Achievement Self-efficacy Scale (MASS): A valid and reliable tool. *Med Teach* 2013;35:575–80. doi:10.3109/0142159X.2013.798401.

[15] Eva KW, Reiter HI, Rosenfeld J, Norman GR. The ability of the multiple mini-interview to predict preclerkship performance in medical school. *Acad Med J Assoc Am Med Coll* 2004;79:S40–42.

Chapter 5. The Effect of Language on the Assessment of Communication Skills Using a Standardised Measurement Instrument in OSCEs

This chapter was submitted for publication in Education for Health, 2015.

5.1. Abstract

Objective. Without a standard measure, the comparison of Communication Skills (CS) outcomes between different settings of an Objective Structured Clinical Examination (OSCE) (i.e. between students, modules, years and institutions) is very challenging if not impossible. We examine the CS type – section 1, section 2, section 3 of the MAAS-Global – that affect students' CS performance in OSCEs, also taking into consideration their first (L1) or second (L2) language.

Method. This retrospective cohort study evaluated the CS components of all OSCE rubrics (i.e. assessment forms) used by 5 different disciplines contributing to three different cohorts of undergraduate medical students. A Two-way ANOVA was used for the comparison of the mean of MAAS-Global score (MG Score) for each group of students and each of the academic terms. G- and D-studies of the Generalizability Theory (G-Theory) were used to calculate the reliability of each OSCE. Significance levels were set at .05.

Results. A generalizability analysis for the 1MB, 2MB, 3MB and 4MB OSCEs revealed that reliability of the OSCEs ranged from $G = 0.28$ to $G = 0.79$ with a median of $G = 0.62$. The reliability analysis of the rubrics calibration for the Disciplines of Medicine, General Practice, Psychiatry, Obstetrics & Gynaecology, and Paediatrics were 0.69, 0.83, 0.99, 0.45, and 0.79 respectively. Only in the Year 4 OSCEs did L1 students perform significantly better than L2 students. This was the case for stations designed by General Practice ($F(1, 368) = 21.46, p < .001$), Psychiatry ($F(1, 368) = 72.94, p < .001$) and Paediatric stations ($F(1, 363) = 8.72, p < .003$). No significant differences were found for stations designed by Obstetrics & Gynaecology.

Conclusion. It is possible to measure MG characteristics and analyse influences between L1 and L2 student's outcome in CS assessment during OSCEs. L2 students perform less well than their L1 colleagues, in some OSCEs. Future research should use a longitudinal design, following students throughout the curriculum and after graduation.

5.2. Background

Objective Structured Clinical Examination (OSCE) can be used to examine and compare the outcome of procedural skills and performance measures, including Communication Skills (CS), between students, modules, years and across institutions, however such comparison is very challenging [1]. One of the central reasons this difficulty is the fact that standardized measurement instruments are rarely used in CS assessment when it comes to formative and/or summative assessment of undergraduate medical students [2]. Whilst the OSCE is the common tool used to assess medical student's clinical and communication skills, according to our previous work, using a standardized instrument is not common practice in the assessment of CS in OSCEs. This problem has been highlighted by similar studies [2–4].

Since the measurement instruments (i.e. scoring rubrics such as station checklist forms) are not standardized in OSCEs, calibration of the instruments is required in order to enable comparison of standardized (i.e. calibrated) results. In a previous study, we proposed the calibration of the existing scoring rubric to assess CS according to the MAAS-Global. The MAAS-Global is an instrument developed by Maastricht University to assess physicians' CS. It has been proven to be valid and reliable [1,5]. According to our proposed calibration procedure we calculate a MAAS-Global (MG) proportion and the MG score [1,6]. The MG proportion is the percentage of the station scoring rubric that maps to CS items based on the MAAS-Global. The MG score, which is the MG proportion times the candidate's raw score of the station, can be considered as the 'standardised' CS

score. This 'standardised' CS score can then be categorized into three types of CS and divided into section 1, section 2, and section 3 items according to the MAAS-Global. Section 1 represents the sequential process of the medical consultation starting from 'introduction' and including 'management' and 'evaluation of the consultation' (this consists of seven items or seven individual processes of the consultation). Section 2 measures the general communication competence that occurs throughout the entire consultation (six items). Whilst section 3 is focused on the quality of the medical content which represents the following phases of the consultation: history taking, physical examination, diagnosis, and management (four items). The clinical skills in the MAAS-Global are only broadly defined because they are case specific and therefore need to be specifically defined for each skill [7].

Nowadays, many medical and health care educational institutions attract students of different nationalities. For many of these students the language of the institution is not their preferred first language. It is evident that medical students may experience difficulties if their native tongue differs from the language used at the educational institution they attend [8,9]. The effect of the language on CS is an important factor in assessing student's CS proficiency and considering their performance in OSCEs. Studies in the area of nursing education confirm that students who are instructed in their first and preferred language (L1) perform better when compared to students where the institutional instructional language is their second or even third language (L2) [9]. However, there is very little existing research evidence about the effect of language on medical student's CS [8,9].

One theory that may help to explain these observed differences is the modern theory of Working Memory (WM). This theory suggests that L2 students are disadvantaged when compared to L1 students. WM is required in cognitive processes and has two functions; the storage of information and information

processing (i.e. analysis of the information) [10,11]. WM consists of a two-tier structure, the first tier is the short-term memory (ST) which is highly accessible but has limited capacity. The second tier is long-term memory (LT) which is not capacity-limited but is not immediately accessible [12]. In L2 students, to analyze retrieved information for the purpose of making a decision, the information needs to be stored in the short-term memory first (9). At the same time, to give meaning to the retrieved information, WM needs to access long-term knowledge which is stored in the long-term memory. In L2 students, the short-term memory is being used to store unfamiliar words, accents and sentences. Therefore when processing information for the main task, the short-term memory is occupied and the result is a decline of the thinking ability [13].

The L2 disadvantage is not attached to a specific language since it can happen in any native language [14]. The WM capacity is limited for those students who use a second language environment for communication and learning [13]. The process of WM in L1 and L2 differs in terms of the type of information it processes. For L2 students mastering the language is not only the ability to understand written and formal language, equally important is the ability to understand the cues and local norms in CS in healthcare. WM processes are even more necessary for L2 students than for L1 students as L2 students need to store information in ST memory for immediate use to enable them to decode unfamiliar language terms and cues [12]. At the same time, WM tries to access LT memory, in order to simultaneously analyze the main OSCE task. The decoding of unfamiliar information is therefore potentially interfering with the main (clinical) task resulting in a possible decrease in the ability to solve and analyze multiple processes in WM. Students with higher language proficiency, experience better conditions compared to those students with a lower language proficiency [10]. L2 students processing the information require more cognitive resources compared to L1 students [12]. Furthermore, potentially disadvantaging factors for L2 students are the greater stress levels due to the non-native environment,

the examination environment and noise. To explain the concept of “noise” Mann stated in 2012: *“Linguistically, English target speech and English speech noise consist of many common properties (e.g., phonemes, syllable structures, prosodic features, etc.), which may make it more difficult for listeners, particularly non-native, to segregate target language from background noise and this may contribute to greater informational masking”* [13]. This will further decrease the WM performance. Whilst in low noise conditions (e.g. most pre-clinical OSCEs), the L2 students should have more similar performances when compared to the L1 students [13,15].

We don't pretend to be able to analyze all factors involved in language proficiency; neither do we pretend to know possible causes of any significant differences between L1 and L2 medical students. However, what we can do is to critically appraise our analysis and possibly determine a pattern of the MG score between L1 and L2 students – including the different sections (1, 2, and 3) of CS (based on MAAS-Global). This is subsequent to standardization of CS scores, as explained earlier. Moreover, the focus of the assessment in each section of MAAS-Global is different [1]. Section 1 and 2 are more about recalling a simple task in the medical consultation process. Whilst in section 3, measuring the content quality in the CS, students have to use higher cognitive processes in order to integrate the information, analyze it, and then make decisions in every phase of the consultation. Therefore, the aim of this study is to examine the CS - in terms of the section 1, section 2, and section 3 types of items according to the MAAS-Global – that affect students' CS performance in OSCEs.

5.3. Method

5.3.1. Context of the Study

This study evaluated the CS components of all OSCE assessment forms used by the Disciplines of Medicine, General Practice, Psychiatry, Obstetrics &

Gynaecology and Paediatrics in the undergraduate medical programme at School of Medicine in the College of Medicine, Nursing & Health Sciences at the National University of Ireland Galway. These disciplines administered OSCEs for years one to four of the undergraduate medical programme using the OSCE Management Information System (OMIS) (Qpercom Ltd <http://www.qpercom.ie>). Ethical approval was granted by the NUI Galway Research Ethic Committee.

5.3.2. Research Subjects

This study analyzed data that was obtained from OMIS from the student cohorts of the 2010/2011, 2011/2012, and 2012/2013 academic terms. The Discipline of Medicine administered the pre-clinical OSCEs for 1MB (first year), 2MB (second year) and clinical OSCEs in 3MB (third year). Four disciplines – General Practice, Psychiatry, Obstetrics & Gynaecology, and Paediatrics – contribute to the clinical OSCEs in 4MB (fourth year). In the 4MB OSCE the total number of stations included a combination of different sets of stations from General Practice, Psychiatry, Obstetrics & Gynaecology and Paediatrics for the two consecutive sessions.

Total number of students (n) for 1MB for the academic terms 2010/2011, 2011/2012, 2012/2013 were 135, 212, and 213 respectively (total n=560). Total number of students (n) for 2MB for the academic terms 2010/2011, 2011/2012, 2012/2013 were 119, 126, and 213 respectively (total n=458). An expansion of the student intake numbers at the medical school took place during these terms. The total number of students for 3MB for the academic terms 2010/2011, 2011/2012, 2012/2013 were 123, 132, and 137 respectively (total n=392). The total number of students for 4MB for the academic terms 2010/2011, 2011/2012, 2012/2013 were 116, 123, and 141 respectively (total n=380).

In this study, students who were citizens of countries where English is the official first language were grouped as L1 students. Students who were citizens of the countries where English is not the official first language, were grouped as L2 students.

5.3.3. Rubrics Calibration

A previous paper describes how the OSCE assessment forms were standardised against the MAAS global to yield a MG proportion and for each student a MG standardised communication score [1].

5.3.4. The Standard

The MAAS-Global consists of 17 items divided into the 3 sections as explained in the background information. Section 1 consists of 7 items, which is a sequence of logical order of consultation. These items are about the first introduction, a follow-up consultation, a request for help, physical examination, diagnosis, management, and evaluation of the consultation. In section 2, the MAAS-Global measures the general CS which occurs in all phases of consultation. The 6 items of section 2 are about exploration, emotions, information giving, summarization, structuring, and empathy. While in section 3, the quality of the medical content during consultation is measured. There are 4 items in this section which are about history taking, physical examination, diagnosis, and management.

5.3.5. Statistical Analysis

A Two-way ANOVA was used for the comparison of the mean of MAAS-Global score (MG Score) for each group of students and each of the academic terms. Following the recommendation of the AMEE Guide 68, for the OSCEs reliability analysis in this study we used the Generalizability Theory [16]. G- and D- studies of the G-Theory were used to calculate the reliability of each OSCE. Significance levels were set at .05. All ANOVA analyses were carried out using SPSS (version 21) and GT were calculated using EduG (Version 6.1-e, by the Swiss Society for

Research in Education Working Group; Edometrics-Quality of Measurement in Education).

5.4. Results

5.4.1. The OSCEs

The number of students assessed in OSCEs during each academic term were as follows: 2010/2011 (n=493), 2011/2012 (n=593) and 2012/2013 (n=704). As shown in Table 5.1, in the 1MB to 4MB OSCEs, the majority of the L2 students originated from Malaysia, whilst the majority of L1 students were from Ireland.

In all OSCEs, the station duration was 5 minutes, with the exception of stations from the Discipline of Psychiatry whereby a 10 minute station duration was employed. Table 5.2 demonstrates that for 1MB and 2MB, each OSCE used five stations. In 3MB and 4MB the majority of OSCEs had ten stations. The exception to this was the Discipline of Psychiatry OSCEs, which had 4 or 5 stations, of longer duration. Whilst the total number of stations varied between the different cohorts and circuits, each student undertook the same duration of assessment (Table 5.2). 4MB students rotate between disciplines over the course of the academic term; hence the OSCE is split over two sittings to accommodate the class. Each student sits the same number of stations and alternative forms of the same OSCE stations at each sitting.

Table 5.1: Summary of L1 and L2 students' country of origin.

Academic Year		n of students		n of students
1MB	L1	403 (72%)	L2	157 (28%)
	Australia	1	Denmark	1
	Canada	40	Finland	1
	UK	6	France	1
	Ireland	331	South Korea	1
	Nigeria	1	Malaysia	151
	Singapore	15	Norway	1
	USA	9	Sweden	1
	2MB	L1	364 (79.5%)	L2

	Australia	1	France	1
	Canada	34	South Korea	1
	UK	7	Malaysia	90
	Ireland	300	Norway	1
	Nigeria	1	Sweden	1
	Singapore	12		
	Trinidad and Tobago	1		
	USA	8		
3MB	L1	320 (81.6%)	L2	72 (18.4%)
	Australia	1	Bangladesh	1
	Canada	26	Indonesia	1
	Ireland	268	Iraq	1
	Singapore	11	Malaysia	67
	Trinidad and Tobago	1	Poland	2
	USA	6		
	UK	6		
	South Africa	1		
4MB	L1	299 (78.7%)	L2	81 (21.3%)
	Canada	15	Indonesia	1
	Gibraltar	1	India	1
	Ireland	261	Iraq	1
	Singapore	8	Malaysia	74
	Trinidad and Tobago	1	Pakistan	1
	USA	3	Poland	2
	UK	9	Sudan	1
	South Africa	1		

A generalizability analysis was performed independently for each of the 1MB, 2MB and 3MB OSCEs. As for 4 MB, analyses were performed separately for each circuit of the OSCEs and for each contributing Discipline. From Table 5.2, it can be seen that the reliability of the OSCEs ranged from $G= 0.28$ to $G= 0.79$ with a median of $G=0.62$. It is apparent that 5 stations in 1MB and 2MB produced lower G-coefficients compared to the 8 or 10 stations in 3MB. In the 4MB OSCEs, the average G-coefficient in the Discipline of General Practice was 0.71 ($sd=0.06$) which was higher than that of the other three disciplines.

5.4.2. Rubrics Calibration

It can be seen in Table 5.3, that the MG proportions (i.e. the percentage of CS tested in each OSCE that can be mapped onto MAAS- Global) varied between 18 and 92%. For those OSCE stations that were administered by the Discipline of Medicine, the average of MG proportion of 1MB, 2MB and 3MB were 66%, 70% and 66% respectively. The MG proportion for 1MB ranged from 64% to 67%, whilst for 2MB this ranged from 66% to 72% and for 3MB ranged from 65% to 67%. It is apparent from Table 5.3 that section 3 of the MAAS-Global (i.e. mastery of the medical content during consultation) was above 50 % in all academic terms. In 3MB the section 3 proportion reached nearly 90%. While for 1MB and 2MB, section 3 proportion ranged from 57% to 69%, with an average of 64.5%.

Table 5.3 also demonstrates that MAAS-Global items 2 (i.e. follow-up consultation) and 7 (i.e. evaluation of consultation) were assessed once in 3MB 2010/2011 and 1MB 2010/11. For all year terms of 1MB, 2MB, and 3MB, item 14 (i.e. history taking) and 15 (i.e. physical examination) of the MAAS-Global were always included in the OSCE rubric, while item 9 of MAAS-Global (i.e. exploring feelings or emotions) was not found in Discipline of Medicine's OSCEs.

Furthermore, Table 5.3 shows that in the year 4 OSCEs the averages of the MG proportion in stations designed by the Discipline of General Practice, Psychiatry, Obstetrics & Gynaecology and Paediatrics were 69%, 79%, 24%, and 50% respectively. The majority of the OSCEs incorporated 50 percent or more of sections 3 of the MAAS-Global, whilst those from the Discipline of Psychiatry incorporated more than 80% of section 3. In stations designed by the Discipline of Obstetrics & Gynaecology, the majority of the rubrics items mapped to section 3 of the MAAS-Global (60% or more). From here onwards, the percentage of the MAAS-Global sections after the calibration process will be referred to as MAAS-Global characteristic (MG characteristic).

5.4.3. MAAS-Global Score

As described in the background, a standardised CM score was calculated for each student by multiplying the MG proportion by the student's raw score. Due to the consecutive administration of OSCEs, each cohort of students was split into two different groups. Each group was assessed in a different environment (i.e. a different group of stations and a different geographical site). According to the AMEE guidelines (AMEE Guide No. 81), these should be considered as different circuits requiring individual circuit analysis [17,18]. Thus, in the 4MB OSCEs, the term "circuit1" will be used to refer OSCEs held in February of each academic term and "circuit2" will be used to refer to those held each April.

The final two columns of Table 5.2 show the summary of the mean MG score of each student group in the academic terms. For 1MB, 2MB and 3MB, analyses were conducted independently. A two-way analysis of variance was conducted on the influence of two independent variables (student groups and academic terms) on the MG score. Student groups consisted of two levels (L1, L2) and academic terms consisted of three levels (2010/2011, 2011/2012, 2012/2013). Despite that the majority of the MG score for L2 students seems to be higher than their counterpart L1 for the results of 1MB, 2MB, and 3MB, no statistical significant differences between L1 and L2 students were found.

Table 5.2: Summary of OSCE settings, Generalizability Coefficient and Decision Study (with 10 and 15 stations), and MAAS-Global score (L1 and L2 students).

Academic year	Academic Terms	Circuit	OSCE contributor	Number of Stations	Station's duration (in minutes)	G	D-Study		MG score (sd)		
							10 stations	15 stations	L1	L2	
1 MB	2010/11	April	Discipline of Medicine	5	5	0.57	0.72	0.80	46.3 (5.2)	48.4 (4.7)	
	2011/12	March		5	5	0.47	0.63	0.72	51.4 (4.1)	52.2 (4)	
	2012/13	March		5	5	0.48	0.65	0.73	53.7 (3.5)	53.2 (4)	
2 MB	2010/11	March	Discipline of Medicine	5	5	0.44	0.61	0.70	52.5 (4.6)	53.7 (4.7)	
	2011/12	March		5	5	0.44	0.61	0.70	50.9 (4.3)	48.6 (3.9)	
	2012/13	March		5	5	0.50	0.67	0.75	58.9 (4.7)	57.9 (3.8)	
3 MB	2010/11	May	Discipline of Medicine	10	5	0.67	–	0.75	46 (3.7)	46.8 (3.4)	
	2011/12	May		10	5	0.62	–	0.71	47.4 (3)	48.4 (3.2)	
	2012/13	May		8	5	0.50	–	0.65	48.5 (3.7)	46.9 (3.3)	
4 MB	2010/11	February	Discipline of General Practice	10	5	0.75	–	0.82	43.9 (4.2)	43.5 (3.5)*	
		April		10	5	0.73	–	0.80	46.1 (3.5)	44.6 (3.3)*	
		February	Discipline of Psychiatry	4	10	0.54	0.74	0.85	65.4 (7.2)	54.2 (8.1)*	
		April		4	10	0.61	0.79	0.85	62.2 (7)	53.4 (8.9)*	
		February	Discipline of Obstetrics & Gynecology	10	5	0.48	–	0.58	31.1 (6.3)	29 (7.1)	
		April		10	5	0.52	–	0.62	30.7 (3.8)	31 (2.9)	
		February	Discipline of Paediatrics	7	5	0.28	–	0.46	59.6 (6.4)	57.8 (5.2)*	
		April		9	5	0.50	–	0.63	54.9 (4.9)	54.6 (4.7)*	
		2011/12	February	Discipline of General Practice	10	5	0.59	–	0.68	44 (3.1)	43.6 (3.7)*
			April		10	5	0.73	–	0.80	46.6 (3.9)	46 (5.3)*

	February	Discipline of	5	10	0.65	0.79	0.85	68.8 (7)	58.9 (6.5)*
	April	Psychiatry	5	10	0.56	0.72	0.79	61.2 (7.3)	57.7 (7.5)*
	February	Discipline of	10	5	0.79	–	0.85	30.8 (4.2)	29 (5.7)
	April	Obstetrics & Gynecology	10	5	0.72	–	0.79	25.2 (3.6)	26.2 (5.8)
	February	Discipline of	10	5	0.61	–	0.71	49.6 (4.1)	47 (6.2)*
	April	Paediatrics	10	5	0.61	–	0.71	47.9 (4.6)	49.5 (4.9)*
2012/13	February	Discipline of General	10	5	0.74	–	0.81	49.4 (3.8)	43 (3.5)*
	April	Practice	10	5	0.70	–	0.77	47.8 (3.9)	43.5 (4.9)*
	February	Discipline of	5	10	0.73	0.84	0.89	61.5 (8.4)	53.3 (6)*
	April	Psychiatry	5	10	0.68	0.81	0.87	60 (8)	53.2 (3.6)*
	February	Discipline of	10	5	0.73	–	0.80	37.9 (3.5)	34.8 (6.3)
	April	Obstetrics & Gynecology	10	5	0.72	–	0.79	39.6 (3.6)	36.8 (6.7)
	February	Discipline of	10	5	0.71	–	0.78	54.2 (4.6)	47.5 (8.2)*
	April	Paediatrics	10	5	0.78	–	0.84	51.3 (4.2)	44.7 (10.6)*

* Significant different

As for the 4MB OSCEs, analyses were conducted independently for each Discipline (Discipline of Obstetrics & Gynaecology, Paediatrics, Psychiatry and General Practice). A two-way analysis of variance was conducted for two independent variables (student groups and circuits) for the MG score in OSCEs. Student groups included two levels (L1, L2) and circuits consisted of six levels (2010/2011 circuit1, 2010/2011 circuit2, 2011/2012 circuit1, 2011/2012 circuit2, 2012/2013 circuit1, 2012/2013 circuit2).

For the Year 4 OSCEs, Table 5.2 demonstrates that L1 students achieved significantly higher results than L2 students in stations designed by the Discipline of General Practice ($F(1, 368) = 21.46, p < .001$). Similar results were found for the stations designed by the Disciplines of Psychiatry and Paediatrics, for which L1 students performed better than L2 students ($F(1, 368) = 72.94, p < .001$) and ($F(1, 363) = 8.72, p < .003$ respectively).

There was a statistically significant difference in interaction between student group and circuits for stations designed by the Disciplines of General Practice and Paediatrics ($p < .05$ level $F(5, 368) = 4.47, p = .001$ and $p < .05$ level $F(5, 363) = 2.49, p = .031$ respectively). As a result, it can be seen in Table 5.2 that in 2011/2012 circuit1 for stations designed by the Discipline of General Practice and 2011/2012 circuit2 for stations designed by the Discipline of Paediatrics, the L1 students performed better than L2 students.

Interestingly, the MG mean score for L1 students has a larger difference compared to the L2 students for stations designed by the Discipline of Psychiatry when compared against stations designed by the Disciplines of General Practice and Paediatrics. There was one exception among the 4MB OSCE stations designed by the Discipline of Obstetrics & Gynaecology, where there were no significant differences found at the .05 level.

Table 5.3: Summary of MAAS-Global proportion, sections and items of stations' checklists in each discipline.

Academic year	Academic Terms	Circuit	OSCE contributor	MG Proportion	Maas-Global (in percent)			Maas-Global Items		
					Section 1	Section 2	Section 3	Section 1	Section 2	Section 3
1 MB	2010/11	April	Discipline of Medicine	64	20	23	57	1,4,5,6,7	10,11,13	14,15,16,17
	2011/12	March		67	21	12	67	1,4,5,6	10,11,13	14.15
	2012/13	March		67	21	20	59	1,3,4,5,6	10,11,13	14.15
2 MB	2010/11	March	Discipline of Medicine	72	14	18	69	1,4,5,6	8,10,11,12,13	14,15,17
	2011/12	March		66	16	15	69	1,3,4,5,6	8,10,11,13	14.15
	2012/13	March		71	19	14	67	1,4,5,6	8,10,11,13	14.15
3 MB	2010/11	May	Discipline of Medicine	65	6	7	87	1,2,4,5	10,11,12,13	14.15
	2011/12	May		65	8	3	89	1,4,5,6	10,11,13	14.15
	2012/13	May		67	6	5	89	1,4,5,6	8,10,11,13	14.15
4 MB	2010/11	February	Discipline of General Practice	75	14	29	57	1,3,4,5,6	8,9,10,11,12,13	14,15,17
		April		69	19	29	52	1,4,5,6	8,9,10,11,12,13	14,15,17
		February	Discipline of Psychiatry	90	13	6	81	5	10	14.17
		April		92	0	0	100	–	–	14,15,17
		February	Discipline of Obstetrics & Gynecology	18	10	28	62	5	10	15
		April		–	–	–	–	–	–	–
		February	Discipline of Paediatrics	48	17	16	67	1,3,4,5	8,10,11,13	14,15,17
		April		54	15	12	73	1,3,4,5,6,	8,10, 8,10,13	14,15,17
	2011/12	February	Discipline of General Practice	65	16	35	49	1,4,5,6	8,9,10,11,12,13	14,15,17
		April		67	11	35	54	1,4,5,6	8,9,10,11,12,13	14,15,17
February		Discipline of	73	2	4	94	5	10	14,15,16,17	

	April	Psychiatry	74	5	6	89	5	10	14,16,17
	February	Discipline of Obstetrics & Gynecology	16	57	27	16	–	10	15
	April	Discipline of Obstetrics & Gynecology	16	0	25	75	5	10	15
	February	Discipline of Paediatrics	40	16	8	76	1,4,5,6	8,10,13	14,15,16,17
	April	Discipline of Paediatrics	51	11	12	77	1,4,5	8,10,13	14,15,16,17
2012/13	February	Discipline of General Practice	66	14	27	59	1,4,5,6	8,9,10,11,12,13	14,15,17
	April	Discipline of General Practice	71	15	30	55	1,3,4,5,6	8,9,10,11,12,13	14,15,17
	February	Discipline of Psychiatry	70	16	7	77	5	10	14,17
	April	Discipline of Psychiatry	73	1	4	95	5	10	14,16,17
	February	Discipline of Obstetrics & Gynecology	33	5	12	83	5	10	14,15,17
	April	Discipline of Obstetrics & Gynecology	36	7	21	72	5	10	14,15,17
	February	Discipline of Paediatrics	54	19	26	55	1,4,5	8,10,11,13	14,15,16,17
	April	Discipline of Paediatrics	53	17	28	55	1,4,5,6	8,10,13	14,15,16,17
	Total		60	14	17	69			

5.5. Discussion

We suggest that the single most important aspect of assessment is the measurement instrument. The instruments used in OSCEs are the station scoring rubrics. A pre-requisite for excellent assessment is the validity and reliability (i.e. reproducibility) of the assessment forms [2,19]. In most OSCE stations a combination of both CS and other clinical skills are assessed in a combined fashion. Knowing this, we endeavored to separate CS items from the other clinical skills items. In the situation where CS instruments were not standardized, a calibration of station scoring rubrics had to be executed in order for the results to become comparable between years of study, participating OSCE contributors and different groups of students (L1 and L2) [1].

Over the years the number of L2 students is increasing internationally [20]. Furthermore, there is an increase in L2 healthcare workers appointed in the healthcare society [21]. As a result of the multi-lingual/cultural backgrounds of our students, we are aware of the potential differences in CS assessment outcomes between different students. Trying to understand the root cause of these differences is critically important in appraising the employed method of CS training and assessment. Analyzing differences between L1 and L2 performance in CS training and assessment is not new, however there is very little research evidence related to the linguistic background of international medical students. To enable a 'fair and honest' (i.e. reliable and valid) comparison between different groups, cohorts and various teaching modules, we calibrated our existing CS assessment forms using the MAAS-Global standard.

The current study found that in all OSCEs administered over a 4 year period, the majority of CS items mapped to sections of the MAAS-Global (Table 5.3), however the contributing Disciplines tended to use different types of MAAS-Global items. OSCE stations developed by the Disciplines of Medicine, General Practice and Paediatrics incorporated a wider range of MAAS-Global items

referring to all three sections of the MAAS-Global. Stations developed by the Disciplines of Psychiatry and Obstetrics & Gynecology focused more specifically on the medical content of the consultation (section 3 of the MAAS-Global) appropriate to their respective disciplines.

The reliability of the calibration of OSCE station rubrics against MAAS-Global was lowest for the Discipline of Obstetrics & Gynecology ($G=0.45$) [1]. A potential explanation for this finding is that the Obstetrics and Gynaecology station rubrics predominantly contained items relevant to clinical skills as opposed to communication skills. According to the user manual MAAS Global does not cater for particular clinical skills [7]. The low reliability of calibration indicates that raters had difficulties on agreeing which Obstetrics and Gynaecology OSCE station rubrics corresponded to MAAS-Global items. Scoring rubric items might need modification to provide this clarity [1]. According to the WM theory, performance of technical / clinical skills would be less impacted by language, hence we would not expect to see a big difference between L1 and L2 if the Obstetrics and Gynaecology stations are indeed more focused on such skills. This was indeed the case in this study.

Building on a previous calibration process, the results of this study provided a 'standardised' CS score, which allows more accurate comparison of results between L1 and L2 students. Research from elsewhere suggests that the CS performance of L2 students is lower when compared to the L1 fellows [10,11,13,22,23]. In contrast to the findings of Mann et al 2013 we did not find any significant difference in performance in the OSCEs in the pre-clinical years [13]. Analysis of results in the 1MB, 2MB and 3MB OSCEs demonstrates that MG scores showed no statistically significant difference between L1 and L2 student groups. A likely explanation is that in these OSCEs students were requested to perform a basic clinical skills assessment, in accordance with the curriculum blueprint. Learning outcomes of 1MB and 2MB concentrated on developing

evidenced-based CS, fundamentals of patient assessment, and developing history-taking and examination skills [24]. Whilst in the 3MB OSCEs, the students start to learn the foundation of patient investigation and diagnosis. In the 4MB OSCEs, students are required to complete the full consultation process from history taking, physical examination, up to analyzing and interpreting findings, and finally synthesis of all available information. As a result of managing the information correctly, they are supposed to formulate a diagnosis and/or differential diagnoses. These processes require a higher level of cognitive ability and availability.

In keeping with this we found statistically significant differences between L1 and L2 students became apparent in the year 4 OSCEs (Table 5.2). We demonstrated that in stations developed by the Disciplines of General Practice, Paediatrics and Psychiatry, L2 students performed significantly worse than L1 students. We suggest that for L2 students, WM was fully occupied with information gathering, language decoding, analyzing and storing unfamiliar sound patterns in their ST memory in addition to trying to activating LT memory to access knowledge processes. The function of WM is that of receiving information, accessing existing knowledge (i.e. that stored in the LT memory), and further analyzing the information to make a decision, plan and implementing the action. In L2 students, WM is so occupied trying to understand the unfamiliar sound patterns, that this interferes with the main task. This results in slower decision making capabilities for L2 students. In addition, it is known that in L2, WM will have reduced decision making capability when there is noise and when stressful tasks are involved [13]. In OSCEs, the 'noise' is caused by the patient communication with the student itself, as well as by the visual information in the station (i.e. patient gesture, cues, health beliefs) and the short period of 5 minutes and even their performance in previous stations. All of these conditions act like a "bottle neck" for WM.

It seems that conceptual thinking and analyzing is very challenging for WM [13]. Advance clinical and communication skills assessment and the high percentage of section 3 items largely occupies the process of WM. These findings may help us to understand that the MG characteristic (i.e. the percentage of section 1, 2 and 3 types of items) is related to the students' performance in OSCEs. The Discipline of Psychiatry demonstrated the largest differences between L2 and L1 groups. In contrast to the other disciplines, Psychiatry stations had the highest MG proportion and incorporated most section 3 items of the MAAS-Global. A possible explanation for this phenomenon might be that in Psychiatry a higher level of CS items is required in a station and more analytical skills, decision making and appropriate action. As a consecutive result, a larger bottleneck condition occurs for L2 students in Psychiatry stations, and hence they perform significantly lower (> 10%) than L1 students. The corollary holds true for the Discipline Obstetrics & Gynaecology, where no significant difference was found between L1 and L2 students. These stations, being largely clinically focused, had the lowest MG proportions. We suggest that this means that the L2 students do not have to decode 'unfamiliar' terms, hence there is no interference with WM with regard to analyzing the information and making a decision.

Good language skills are not only a matter of mastering reading and writing. Students must also be able to interpret language in the context of non-verbal and cultural values [12]. While CS are heavily connected to the context and content of the clinical aspect, understanding the cues in a consultation process means the student needs to understand also the social norms, health and personal beliefs of the patient [25]. The L2 student presumably does not have the privilege of taking in and processing that type of information, which is unique to the social environment where the higher education institution is located. In this case, processing the unfamiliar information will occupy the WM completely. The findings of this study support previous studies that non-native speaking students demonstrate weaker performance when compared to the native

speakers due to a language disadvantage. The question however is, what aspects of the blueprint in the Educational Institution highlights the educational goals for L2 students integration to enhance their internships and daily practice where they will have to communicate with mostly L1 patients? Assessment and training in English for Medical Purposes and English for Specific Purposes by professionals has been recommended elsewhere [26]. Conversational language and cultural awareness training is offered to international medical students at our institution, but it is not compulsory and the uptake is variable. Consideration of how to best encourage students to see the value of and promote engagement with these and other measures, such as early immersion programmes, is warranted.

With an average statistical significant lower performance in Year 4 of the curriculum for L2 students of around 10% (SD) one might question whether this statistical difference has a clinical (i.e. educational) impact. Information gathering is similar between L1 and L2 students in the years 1, 2 and 3 however when it comes to analyzing the relevant information, clinical decision making and treatment planning, the L2 students require more WM to perform in what is for them a more stressful (non-native) environment. However, to discuss whether a statistical significant difference of 10% has any 'clinical or educational' impact, follow up research in clinical education (i.e. on patient satisfaction and impact on decision making) is required.

Turning now to the OSCEs reliability, according to Cardinet et al. (2012) an assessment is consider reliable if the G-coefficient is ≥ 0.80 [27]. In this study, the reliability of the OSCEs range from fair to good, based on the G-theory analysis. It can be seen that OSCEs with 4 and 5 station never reached an acceptable G-coefficient. However, OSCEs that were administered by the Discipline of Psychiatry – during the 2012-2013 academic term – achieved a fairly high G-coefficient. This result is plausible considering the length of the station in the Discipline of Psychiatry's OSCE were 10 minutes compared to the Discipline of

Medicine's OSCE, which were only 5 minutes long. In the literature, station duration has been associated with the reliability of an assessment [28]. According to the corresponding D-study, increasing the number of stations to at least 10 or 15 stations will generate higher G-coefficients and therefore improved reliability. This is consistent with previous studies that a greater amount of stations can produce a higher reliability [19]. In the D-Study of the 1MB and 2MB cohorts, a hypothetical amount of a 10 station OSCE would achieve approximately the same G-coefficient as compared to the 3MB OSCE using the actual number of 10 stations. With the hypothetical amount of 15 stations, the reliability coefficient of the OSCEs still ranges from a lowest G coefficient of 0.46 to a G coefficient of 0.89 with a median of 0.78. These findings are consistent with earlier studies which have generally indicated that to produce a higher reliability coefficient a higher number of OSCE stations is required [19]. Increasing the number of the stations, as demonstrated in the D-study, will help to increase the reliability of the OSCEs, and therefore yield a more robust outcome [18,19]. Another option that could be considered to improve the reliability of the OSCE is to add a second rater to each station [19]. In this study, the findings revealed that in several OSCEs with 10 stations reliability was unsatisfactory. Despite using the G-theory analysis the cause of this unanticipated finding remains unknown and therefore further study is required using better designed and an improved number of stations using multiple examiners [19]. The OMIS software does support a multiple-examiners feature and, to avoid further increasing of the running cost of an OSCE, the Standardized Patient (SP) could be added as an extra reliable rater [27].

5.6. Limitations of the Study

The results of this study may be limited because we did not carry out a consecutive analysis of one cohort of L1 and L2 students during the full course of the 5 year curriculum. Substantial changes to the medical curriculum and the

lack of electronic marking of OSCEs for the final year of the programme precluded us from doing so. The effect of the MG characteristics could be further explored if we could follow up students over a longer period of time (e.g. throughout the entire curriculum of training). This would allow not only the assessment of 'change over time', but also examination of whether change is due to improvements in CS training or 'error around the observed score'. Therefore further study is required to follow up the sample from the last 2 academic years 2013/2014 and 2014/2015 to have a better understanding of MG characteristics in one cohort of the students over the course of their CS training curriculum.

This study did not incorporate the entry level language proficiency test results of the L2 students nor their years of training in English. Moreover, the difference between male and female students needs to be assessed. All of these factors are important and require further analysis and consideration.

5.7. Conclusion

We demonstrated that the calibration of OSCE scoring rubrics using the MAAS-Global standard provides detailed information about the MG characteristics of each OSCE. Working memory theory has provided a frame work to attempt to understand the differences that emerge between L1 and L2 students' OSCE CS scores, as they progress through the medical curriculum and are assessed in increasingly complex OSCE scenarios. Future research should explore the effect of MG characteristics on students' performance in CS, using a longitudinal design, which follows cohorts of students throughout the curriculum and even after graduation. Future theory driven research using Working Memory should examine if providing additional time in the OSCE to L2 students would result in improved OSCE CS performance.

5.8. References

- [1] W. Setyonugroho, T. Kropmans, K.M. Kennedy, B. Stewart, J. van Dalen, Calibration of Communication Skills Items in OSCE Checklists according to the

- MAAS-Global., Patient Educ. Couns. (n.d.). doi:10.1016/j.pec.2015.08.001.
- [2] W. Setyonugroho, K.M. Kennedy, T.J.B. Kropmans, Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review, Patient Educ. Couns. (2015). doi:10.1016/j.pec.2015.06.004.
- [3] R.S. Beck, R. Daughtridge, P.D. Sloane, Physician-patient communication in the primary care office: a systematic review., J. Am. Board Fam. Pract. 15 (2002) 25–38.
- [4] H. Boon, M. Stewart, Patient-physician communication assessment instruments:: 1986 to 1996 in review, Patient Educ. Couns. 35 (1998) 161–176. doi:10.1016/S0738-3991(98)00063-9.
- [5] J.M. van Es, C.J.W. Schrijver, R.H.H. Oberink, M.R.M. Visser, Two-dimensional structure of the MAAS-Global rating list for consultation skills of doctors, Med. Teach. (2012). doi:10.3109/0142159X.2012.709652.
- [6] W. Setyonugroho, thomas Kropmans, P. Hayes, R. Murphy, J. van Dalen, K.M. Kennedy, True Communication Skills Assessment in Interdepartmental OSCE Stations: Comparison of Multiple Instruments, (2015). Manuscript submitted for publication.
- [7] Van Thiel J, Ram P, van Dalen J. MAAS-global manual. Maastricht Maastricht Univ 2000:4–5. (http://www.hag.unimaas.nl/Maas-global_2000/GB/MAAS-Global-2000-EN.pdf): (accessed 8.11.2012).
- [8] T.T. Donnelly, E. McKiel, J.J. Hwang, Challenges and motivators influencing the academic performance of English as an additional language (EAL) nursing students: the perspectives of the students, CJNR Can. J. Nurs. Res. 41 (2009) 130–150.
- [9] T. Harvey, C. Robinson, R. Frohman, Preparing culturally and linguistically diverse nursing students for clinical practice in the health care setting, J. Nurs. Educ. 52 (2013) 365–370. doi:10.3928/01484834-20130529-02.
- [10] J. Kormos, A. SáFáR, Phonological short-term memory, working memory and foreign language performance in intensive language learning, Biling. Lang. Cogn. 11 (2008). doi:10.1017/S1366728908003416.
- [11] K. Hye, R.A. Sevcik, The role of verbal working memory in second language reading fluency and comprehension: A comparison of English and Korean, Int. Electron. J. Elem. Educ. (2011) 47.
- [12] J.A. Linck, P. Osthus, J.T. Koeth, M.F. Bunting, Working memory and second language comprehension and production: A meta-analysis, Psychon. Bull. Rev. 21 (2014) 861–883. doi:10.3758/s13423-013-0565-2.
- [13] C. Mann, B.J. Canny, D.H. Reser, R. Rajan, Poorer verbal working memory for a second language selectively impacts academic achievement in university medical students, PeerJ. 1 (2013) e22. doi:10.7717/peerj.22.
- [14] C. Alptekin, G. Erçetin, The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading, J. Res. Read. 33 (2010) 206–219. doi:10.1111/j.1467-9817.2009.01412.x.

- [15] J. Zhang, L. Xie, Y. Li, M. Chatterjee, N. Ding, How Noise and Language Proficiency Influence Speech Recognition by Individual Non-Native Listeners, *PLoS ONE*. 9 (2014) e113386. doi:10.1371/journal.pone.0113386.
- [16] R. Bloch, G. Norman, Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68, *Med. Teach.* 34 (2012) 960–992. doi:10.3109/0142159X.2012.703791.
- [17] K.Z. Khan, K. Gaunt, S. Ramachandran, P. Pushkar, The Objective Structured Clinical Examination (OSCE): AMEE Guide No. 81. Part II: Organisation & Administration, *Med. Teach.* 35 (2013) e1447–e1463. doi:10.3109/0142159X.2013.818635.
- [18] T. Swanwick, *Understanding Medical Education: Evidence, Theory and Practice*, John Wiley & Sons, 2013.
- [19] M.T. Brannick, H.T. Erol-Korkmaz, M. Prewett, A systematic review of the reliability of objective structured clinical examination scores, *Med. Educ.* 45 (2011) 1181–1189. doi:10.1111/j.1365-2923.2011.04075.x.
- [20] J.A. Hallock, D.W. McKinley, J.R. Boulet, Migration of doctors for undergraduate medical education, *Med. Teach.* 29 (2007) 98–105. doi:10.1080/01421590701268723.
- [21] H.M. Part, R.J. Markert, Predicting the first-year performances of international medical graduates in an internal medicine residency, *Acad. Med. J. Assoc. Am. Med. Coll.* 68 (1993) 856–858.
- [22] A. Fernandez, F. Wang, M. Braveman, L.K. Finkas, K.E. Hauer, Impact of Student Ethnicity and Primary Childhood Language on Communication Skill Assessment in a Clinical Performance Examination, *J. Gen. Intern. Med.* 22 (2007) 1155–1160. doi:10.1007/s11606-007-0250-0.
- [23] R.B. Hays, P. Pearse, C.W. Cooper, L. Sanderson, Language background and communication skills of medical students., *Ethn. Health.* 1 (1996) 383–388.
- [24] School of Medicine National University of Ireland Galway, *The Undergraduate Medical Curriculum MB. BCh. B.A.O degree*, (2014).
- [25] T.T. Donnelly, E. McKiel, J. Hwang, Factors influencing the performance of English as an Additional Language nursing students: instructors' perspectives, *Nurs. Inq.* 16 (2009) 201–211.
- [26] S. Eggly, J. Musial, J. Smulowitz, The Relationship Between English Language Proficiency and Success as a Medical Resident., *Engl. Specif. Purp.* 18 (1999) 201–208.
- [27] Cardinet J, Johnson S, Pini G. *Applying Generalizability Theory Using Edug*. Taylor & Francis; 2012.
- [28] D.I. Newble, D.B. Swanson, Psychometric characteristics of the objective structured clinical examination, *Medical Education.* 22 (1988) 325–334. doi:10.1111/j.1365-2923.1988.tb00761.x.
- [29] S. Kilminster, T. Roberts, P. Morris, Incorporating patients' assessments into objective structured clinical examinations, *Educ. Health Abingdon Engl.* 20 (2007) 6.

Chapter 6. General Conclusion

6.1. Preface

This thesis had the goal of understanding and finding a 'gold standard' measurement instrument of CS within the medical education environment. A particular area of focus concerned undergraduate medical students and the use of OSCE as the assessment tool. Furthermore, determining what is used as the 'gold standard' in the learning process was also of interest. In total, four studies investigated the assessment of CS in undergraduate medical students. The result of this thesis is expected to improve the learning process of the students in specific areas by building better assessment tools to assess CS. This thesis investigates whether there are applicable instruments for assessing CS which could be valid and reliable, making CS students' competence assessment comparable, and giving more insight into existing assessments based on a 'gold standard'. The following research questions were formulated:

1. Is there any agreement amongst researchers and medical educators on a measurement instrument to assess the CS of undergraduate medical students in OSCEs? (Chapter 2)
2. What is the method used to calibrate the existing CS assessment of the OSCEs, which have different standards? (Chapter 3)
3. Are there any differences in CS assessed from different settings of OSCEs? (Chapter 4)
4. Is there any relationship between the type of CS affecting the performance of students when studying in a foreign language environment (L2)? (Chapter 5)
5. How many OSCE stations will produce a reliable examination? (Chapter 5)

The results of these studies are summarised, and, by using the evidence presented in each chapter, the implications from the findings can be used to further improve the CS assessment in our and other medical schools in regard to the use of a standardised instrument, and to calibrate existing instruments. This

chapter also presents a conclusion and recommendations, with the latter being useful for future work.

6.2. General Discussion

This thesis has provided evidence that there is still no agreement amongst researchers and medical educators as regards a measurement instrument to assess CS in OSCEs. A method of calibrating the existing rubrics in OSCEs was created. This calibration has been applied to existing data that were managed by our OSCE Management Information System (OMIS), and the findings were promising. Further sections will explain the findings of this thesis, including examples of the practical use of the calibration of existing OSCE rubrics.

6.2.1. A standard of Communication Skills (CS) Measurement Instrument

A reliable measurement instrument used to assess the CS of undergraduate medical students is important, in order to ensure that the assessment produces fewer errors, and the result thus be meaningfully interpreted [1,2]. In Chapter 2, a systematic review was conducted as a background theory and literature review for this thesis. That chapter provided evidence of common practice in CS assessment in OSCEs. The systematic review conducted searches for studies from January 1975 to December 2012 in four databases. 1,998 papers were identified, and after applying several steps of inclusion and exclusion, 34 papers were finally analysed. According to the analysis, 27 domains of CS have been measured in OSCEs in undergraduate medical school programmes. These 27 domains did address two types of assessors: the examiner and the standardised or simulated patient (SP). The CS were focused upon generic CS, doctor-patient communication, history-taking (including focused history-taking), negotiating plan/treatment, information-giving and empathy. Where SP raters were involved as assessors, they focused mainly on generic CS and interpersonal skills.

It is interesting to note that three raters involved in this study encountered difficulties in distinguishing what domain of CS was measured in a particular paper. This study also found that there was an absence of consensus in regard to the interpretation and definition of domains of CS. Only one 'real' measurement instrument was found to be used in CS assessment in OSCEs, the MAAS-Global [3,4]. Three papers indicated that they were using the Calgary-Cambridge Observation Guide (CCOG) as the standard, while, in fact, the creator of CCOG himself mentioned that CCOG was not intended to be used as a measurement instrument [5–8]. In 2014, Burt et.al developed the Global Consultation Rating Scale (GCRS), based on the Calgary-Cambridge guide [9]. This instrument's aim is to assess CS only and differs from MAAS-Global which also measures medical content related to CS (MAAS-Global section 3). The GCRS consists of 12 global areas which measure the CS process excluding its content. Although GCRS shows good reliability in terms of the classical psychometric analysis ($r > 0.80$), it is still lacking of published psychometric evidence in other studies and in terms of the Generalizability Theory analysis. This absence of standardisation precludes the comparison of outcomes across assessment settings. Therefore, a 'real' measurement instrument is suggested for use in terms of CS assessment in OSCEs. Again, studies have been warning about this absence of an agreement since the late 1990s [10,11]. One important consideration regarding the measurement instrument is that re-creating new assessment instruments is not necessary, as existing instruments can be modified to fit different needs.

6.2.2. Calibration Method and Procedures

How do we compare the height of a table and a chair? It would be a simple task, involving the use of a measuring tape. Using either inches or centimetres would be comparable, since both units of length are standardised, and can be converted to the other unit. This could not be achieved if we use a rubber band to measure the length, since rubber is pliable, and the length is not fixed,

meaning that it is not standardised. To relate this to the current scenario, in order to compare the CS of undergraduate medical schools to different cohorts of students, OSCE settings, or even different institutions, a comparable measurement instrument is needed. Therefore, as detailed in Chapter 3, the study was conducted with the aim of seeking a method of calibrating existing measurement instruments. As concluded in Chapter 1, findings from two studies indicated that MAAS-Global is the 'real' measurement instrument. Hence, this study explored the possibilities of calibrating existing rubrics using the MAAS-Global as standard.

MAAS-Global global was developed by Maastricht University to assess physicians' CS [12]. This instrument is divided into three sections, with 17 items in total, and with each item in Section 1 and 2 having sub-items. Section 1 measures the CS for each phase in a medical consultation (seven items). Section 2 contains six items, and is concerned with assessing the general CS that occurs in several or all phases of a consultation. These two sections basically assess whether a doctor did or did not use the CS according to the item list in MAAS-Global. Section 3 is intended to assess the quality of the content related to medical aspects throughout the consultation.

All Year 4 OSCE checklists (n=280) from four academic terms – 2009/2010, 2010/2011, 2011/2012 and 2012/2013 – were analysed by three raters. The penultimate-year OSCEs were contributed by four disciplines: Obstetrics & Gynaecology, Paediatrics, Psychiatry and General Practice. This calibration procedure followed the simple principle of matching each of the OSCE checklist items with the MAAS-Global items. At the end of the process, calculations were made to count the percentage of the checklist items that were considered to be CS according to the MAAS-Global. A calibration manual (MAAS-Global Checklists Calibration Manual) was developed and can be found in the appendix.

This study found that the process of calibration was reliable, as shown by the high G-kappa coefficient ($G=0.80$) [1]. Furthermore, in the case of Decision Study (D-study), the calculation of the hypothetical design using only two raters was considered high ($G = 0.72$). Therefore, we suggested that the calibration of existing OSCE checklists should be conducted by two raters. This would save a great deal of manpower.

Following the calibration of the existing OSCE checklists, this study found interesting results. Nearly half of the items were considered as belonging to Section three of the MAAS-Global, while approximately one third of the items were not considered as CS according to MAAS-Global. The remainders were belonging to Sections 1 and 2 of the MAAS-Global. Chapter 3 also demonstrated that, in CS assessment, the disciplines of General Practice and Paediatrics incorporated the majority of MAAS-Global items in their stations, while the discipline of Obstetrics & Gynaecology and Psychiatry focused mainly upon Section 3 type of items e.g. the quality of the medical content being assessed. Evidence was provided in Chapter 3 that the calibration of existing OSCE checklists is well possible. Moreover, it was demonstrated that a comparison of CS assessments from different settings of OSCEs using a single standard measurement instrument provides comparable information regarding existing CS assessments. Chapter 3 focused on a method of OSCE checklist calibration, using the MAAS-Global as the standard. The following section will discuss further steps after having calibrated OSCE station checklists for CS assessment.

6.2.3. Standardised Communication Skills

The present section describes an application after having the CS assessment calibrated using MAAS-Global as a standard. Whilst previous chapters explained the comparison of the CS assessment of the OSCEs, Chapter 4 provided evidence of the comparison of OSCE outcomes specifically in CS. Two disciplines, General Practice and Psychiatry, partook in this study to analyse students' performance in

OSCE across three academic terms (2009/2010, 2010/2011, 2011/2012 and 2012/2013). With the method presented in Chapter 3, the study identified the percentage of CS in OSCE checklists according to MAAS-Global. Henceforth, this will be referred to as the 'MAAS-Global proportion' (or 'MG proportion'). The study, as detailed in Chapter 4, calculated the student score of each station by multiplying it by the MG proportion of that station. The result was the MAAS-Global score (MG score). In other words, this MG score originated from the CS score that was derived from the OSCE checklist calibration, and this score is considered to be a 'standardised' CS score.

Regarding the MG proportion, two interesting findings were gathered by knowing the MG proportion from each OSCE station designed by faculty of each discipline involved in the OSCE. One is that the higher the MG proportion, the more CS is being assessed in OSCEs. The Psychiatry stations had higher MG proportions than the General Practice stations in each examination. The Psychiatry stations predominantly assessed CS (nearly 80% or MG proportion = 80). Meanwhile, for General Practice stations, this study found that 31% of the checklist items were considered to belong to technical skills assessment, with an MG proportion = 69. Secondly, within the MG proportion, we refer to the percentage of section 1, 2 and 3 type of items of the MAAS-Global and being incorporated in each OSCE. The percentage of each of the MAAS-Global sections will be referred to as to MAAS-Global characteristics (MG characteristics). The MG characteristics were definitely unique to each participating discipline. In General Practice stations, we found that the majority of all types of MAAS-Global items were incorporated in their stations. Meanwhile, in Psychiatry stations, the majority of the items belonged to only Section 3 of the MAAS-Global. With this high percentage of items belonging to Section 3, it could be said that the Psychiatry OSCE stations place an emphasis on assessing the medical content of the consultation only.

This study's results illustrate that the average MG score of the students in Psychiatry stations were higher than in General Practice stations. This phenomenon is a direct result of the fact that Psychiatry stations having a higher MG proportion. Having a 'standardised' CS score is important in order to be able to measure the change and the progress achieved by the students over the period from entering the course to graduating.

Sensitivity in measuring the change is an important aspect of the measurement instrument in psychometric analysis. Further study is required to explore the pass mark for CS based on the calibration process. As already mentioned above, a reliable assessment is needed in order to gain a result which can be meaningfully interpreted. Using the 'real' measurement instrument to assess CS is a must [8]. At the very least, calibrating the OSCE checklists with the known standard, such as MAAS-Global, will be the minimum action.

The two most important findings that have a direct implication on clinical skills assessment in this chapter are MG proportion and MG score. As described in Chapter 3, and further in Chapter 4, each discipline always uses similar patterns across each examination of the MG proportion (including MG characteristics) while designing their OSCE rubrics. As each discipline has its own specific learning outcomes, in clinical skills assessment each discipline needs specific measurement instruments to assess the student in OSCEs. Furthermore, it is interesting that, even before this study proposed the calibration method, each discipline was using an 'unstandardised' measurement instrument which apparently shows a unique assessment pattern for each department. It requires further discussion with participating disciplines as to whether this occurred coincidentally. In developing measurement instruments – i.e. OSCE checklists – to assess students' CS, a long-term study should be conducted in order to ensure that the instruments are valid and reliable and assessment procedures are quality assured. As authors of previous studies have been suggesting since 1997,

a lack of agreement in terms of the standard of CS assessment should not occur in the future [10,11,13]. Researchers and medical educators must avoid 're-inventing the wheel'. Agreement in regard to a gold standard of measurement instrument will have more benefit compared to using locally developed measurement instruments which do not apply universally. A simple illustration of current conditions is that each institution makes their own scale of measurement tape while there are already available standardised measurement tapes (e.g. in metres, inches, feet, etc). Thus, the question to be asked of medical educators is as follows: 'Have we measured what we want to measure? Do we need to create a new instrument to measure?'

In addition, it is important to mention standard settings. There is no strong evidence as regards the origin of the pass mark, for example 50% or 60%. It is worth considering the introduction of criteria for CS with the standardised instrument, or calibrated checklists for the pass criteria, as an addition for the general clinical skills score in OSCEs. This study provides a 'standardised' CS score that is applicable to almost any setting of OSCE. There are two additional benefits from applying standardised CS assessment, for students and for educational institutions. Firstly, students witness their progress throughout the learning process, recognize their weaknesses at every level of training, and are able to engage in self-reflection to improve their skills. Meanwhile, educational institutions can more easily compare the different cohorts of students, quickly identify problems and evaluate the learning process. Although this study provides evidence to calculate the 'standardised' CS score out of the results of different measurement instruments, further research, using a larger sample, will be required in order to apply the method from this study.

6.2.4. MAAS-Global Characteristics and the Second Language (L2)

Many institutions of higher education host a large proportion of international students. As a result, the language of instruction is often not their first language

(L1). Hence, these students inhabit a learning environment in which they must speak and learn through a second language (L2) [14,15]. It has been indicated in previous studies that L2 students are at a disadvantage compared to native students, and this is known to be a factor that influences their performance [15,16]. Therefore, as detailed in Chapter 5, a study was conducted with the aim of examining the effect of MG characteristics in the performance of students who used a second language in the learning process, specifically in relation to the undergraduate programme of a School of Medicine. It is important to note that the disadvantage of L2 students is not attached to a specific language, as this may apply to any native language [17]. On the contrary, this is a common problem whenever students have to study in a learning environment where the language of instruction is not their first language.

Mastering a language does not only involve the ability to understand formal and written language; it also involves knowing local norms, while non-verbal communication is also important [18]. In the clinical environment, CS is highly content and context-based [19]. The ability to understand what the patient seeks to express, and to provide information in a clinical consultation, requires more than formal language. Engaging and maintaining a connection must include non-verbal communication that students must also learn. While L2 students have good language proficiency, they develop more passive language skills (i.e. reading and writing). One study revealed that there is no correlation between the clinical skills of medical students' performance and the language proficiency test unless another variable was added, e.g. spoken language, primary childhood language, language background, etc [20,21].

According to the Working Memory (WM) model, a decision is made after information is received, saving important information in short-term (ST) memory while, at the same time, accessing the long-term (LT) memory and then analysing all of this information. L2 students process differently. They need to understand

and decode unfamiliar information (i.e. language terms, cues). The process of decoding unfamiliar information interferes with the main task, which is to make a decision and subsequently take an action. Hence, an L2 learner takes longer to make a decision than an L1 student. It is also known that in L2, the WM will reduce the capability of making a decision when there is noise and when stressful tasks are involved. Whilst OSCE is found to be more stressful compared to other assessment tools (e.g. written tests) [22]. However, in a quiet and less stressful environment, the L2 students should perform similarly to L1 students [23,24].

Of all the OSCEs from Year One to Four, L1 only performed significantly different to L2 students in Year Four. There were no significant differences between L1 and L2 students, except in the Obstetrics & Gynaecology stations of Year Four. In penultimate-year OSCEs, Psychiatry stations have the highest percentage of items belonging to of Section 3 of the MAAS-Global compared to General Practice and Paediatrics stations. The result shows that in Psychiatry stations L1 students perform better than L2 students; the difference between the two groups is the largest compared to the other two disciplines. Section 3, with an emphasis on assessing the quality of medical content, certainly demanded more cognitive processing compared to items belonging to Section 1 and 2, which involved simply recalling memory and repetitive tasks. In short, the more complex CS to be exercised by L2 students, the slower they would be processed by the brain, thus reducing the capability of processing the main task, because of being occupied with decoding unfamiliar information.

The present findings seem to be consistent with other research, which found that language does have an influence on students' performance. Specifically in CS, a high degree of language proficiency is not enough. Mastering non-verbal communication and local norms or customs are also important. Therefore, educational institutions should consider a kind of programmes that allow the L2

students to engage in social activities which immerse them in the community and force them to leave their comfort zone by communicating in the native language of the L1.

6.2.5. The Reliability of the OSCEs

The most important aspect of all is the reliability of the assessment. As described in Chapter 5, this study analyses all OSCEs from Year One to Year Four across three academic terms (2009/2010, 2010/2011, 2011/2012 and 2012/2013). All OSCEs were analysed using Generalizability Theory (G-Theory), with the aid of the software package EduG. The G-Theory is a reliability analysis that has been recommended by the Association of Medical Education in Europe (AMEE), although, to date, many studies still report reliability analysis using Cronbach Alpha [24,25]. The reliability of the assessment is important because, according to classical test theory, an observation score from the result of an assessment is comprised of two factors, true score and error. The reliability of the assessment basically measures the extent of the stability of the measurement at different times for the same students using the same measurement instrument [26].

With the OSCE as the assessment tool, the main task given to the students was to perform a task in each station. Ideally, the result should not be influenced by external factors (e.g. the examiner or the day of the examination). It should be able to tap into the 'real' skills of the students that have to be mastered during the learning period. A medical educator has the obligatory task of identifying the source of error in an assessment. Therefore, future assessments will have fewer errors. An advantage of using G-Theory for reliability analysis is the ability to ascertain more than one source of error at once and to quantify the error. Using Decision Study (D-study), a hypothetical calculation of the source of variance can also be used to create a simulation of what needs to be changed in the source of variance for future assessments, in order to achieve better reliability [1].

The findings of Chapter 5 were aligned with previous studies, which indicated that a low number of stations in OSCEs will produce low reliability [25]. This can be seen in all OSCEs with five stations, which indicated lower reliability than the accepted standard of reliable assessment (0.80) [1]. The D-study analysis indicated that increasing the number of stations to 10 increased the reliability. Meanwhile, adding more stations meant increasing the cost; adding a second examiner for each station could be the preferred method of achieving higher reliability.

6.3. Generalisation

The studies detailed in this thesis related to OSCEs in the School of Medicine, National University of Ireland Galway. Generalisation of the result from these studies to other schools of medicine or to health education in other countries is possible. Meanwhile, the calibration of OSCE checklists in different schools may be challenging, unless those schools are using electronic administration software for managing the OSCEs. In our school of medicine, a software package called OSCE Management Information System (OMIS) has been used since 2009; therefore, analysis of previous OSCEs is easy (i.e. when obtaining all of the data, including the rubrics). Furthermore, the author is aware that every OSCE is different, and still believes that the calibration of OSCE checklists is possible and required. Another school may still decide to calibrate by using our method, but with different measurement instruments as described in Chapter 3. It is important to remember that it must be a 'real' measurement instrument that has been proven to be valid and reliable. Moreover, it is important to state that every OSCE should be properly evaluated to ensure that it produces reliable assessments.

6.4. Implications for Further Research

CS involves important skills that need to be mastered by the students. A series of studies in this project has been conducted to overcome the problem of the

inability to compare the CS of medical students across different settings of assessments, different cohorts of students, or different schools of medicine. Therefore a number of suggestions have been made throughout the thesis to improve this research, including the following.

This project began with a systematic review. Since the study was limited to measurement instruments in the OSCE in undergraduate medical students, many communication skill measurement instruments were not included in the results of this study [10,11]. However, the findings highlight the same point, which is the lack of agreement regarding the measurement instruments amongst researchers and medical educators. Furthermore, this study found only one 'real' measurement instrument being used by examiners in OSCEs to assess students' CS, which is the MAAS-Global. The next studies in this project were based on these findings. For future study, it is suggested to assess all studies using existing CS measurement instruments, not only at the undergraduate but also at the postgraduate level. Hence, the appropriate 'top down' alignment of CS training and assessment will be adequately reviewed.

- The study attempted to calibrate the existing OSCE station checklists, which calibration process is proven to be reliable. Therefore, it is suggested that medical educators conduct similar studies to calibrate their existing OSCE checklists. However, the calibration process does not need three raters; instead, two raters will produce reliable calibration. Reducing the number of raters will encourage researchers to conduct similar projects, since manpower and time-consumption will also be reduced, and it is crucial to maintain these at low levels if researchers are to deem such research worthwhile.
- After calibrating the checklists, determining the 'standardised' CS score from the OSCE score is another labour-intensive process. Again, our system does not have an automatic process to calibrate and then

calculate the CS score based on standardised checklists. It needs further development of the software in order to ensure an easier process to obtain a 'standardised' CS score. Meanwhile, further study is required to investigate the next step after determining the CS score from standardised checklists. Incorporating it into the standard setting of OSCE by introducing it as another criterion for the pass mark is a subject which requires in-depth study and discussion.

- Having discussed the problems and suggested future research regarding the calibration process and calculating the CS score, the next section addresses the condition whereby the L2 students – students who study in a second language – have a disadvantage in regard to CS. Whilst the finding suggested that the WM theory is likely to be related to the MG characteristics, future study should explore the effect of MG characteristics using longitudinal design (follows cohorts of students throughout the curriculum). Furthermore, for medical educators, there is a question in relation as to whether there are adequate policies in place that could limit the disadvantage of non-native students. A simple suggestion, based on previous studies, would be to encourage the students to engage in social activities or extra class activities that are already available within the university.
- In addition, more research is needed to investigate whether the direct application of MAAS-Global into OSCE, rather than a subsequent calibration, is required. Since MAAS-Global has been proven to be a valid and reliable measurement instrument, creating a new measurement instrument e.g. station forms is not suggested. As also mentioned in Chapter 5, a reliable assessment is very important in order to ensure meaningful results, and further research is necessary with an improved number of stations to analyse each OSCE thoroughly in order to increase

the quality of the assessment and to reduce the number of errors in the results.

6.5. Recommendation for CS Assessment in OSCEs

First and foremost, the most important aspects of an assessment, especially OSCE, are the validity and reliability. As mentioned in previous research, a reliable assessment allows us to gather insightful meaning from the result. The results from analysing all OSCE in three academic terms (2010/2011, 2011/2012/2012/20113) have shown that five stations were an insufficient number to achieve strong reliability. Therefore increasing the number of stations, or adding a rater to each station, could theoretically increase the reliability of the OSCE.

Secondly, regarding CS assessment in OSCEs, it is recommended to use a 'real' measurement instrument. It was already mentioned in a previous study that assessing CS is more difficult than assessing clinical skills [26]. Therefore, using unstandardised instruments is not recommended, as standardised instruments are available. Using a calibration method to calibrate the instrument is suggested whenever the instrument used in OSCE is uniquely developed by the participating disciplines for special purposes.

Thirdly, the 'standardised' CS score revealed that there were differences between L1 and L2 students. This result is aligned with the previous study's findings, namely that language could affect students' performance. Hence, a programme needs to be developed in order to increase the language proficiency of L2 students, not only in formal and passive language, but most importantly in understanding the non-formal language and the local culture in the educational institution.

6.6. Conclusion

This project was undertaken to resolve a 'forgotten' problem, which was pinpointed by Boon and Stewart in 1998. This concerns the lack of agreement

amongst researchers and medical educators on a uniform CS measurement instrument to assess undergraduate medical students [10,11,13]. The use of different measurement instruments means that comparing the results of assessments is challenging and almost impossible to carry out immediately, lacking the convenience of scores such as IELTS and TOEFL, which are immediately comparable across different contexts.

The author conducted a series of studies in this thesis, and the results provide evidence that a comparison of assessment results (i.e. OSCEs) that used different measurement instruments is possible using the calibration method. This calibration involved using an existing measurement instrument, the MAAS-Global, as a standard. Whilst this method is also able to produce a 'standardised' or 'true' CS score, more studies and participating educational institutions are required. The question arises as to whether the CS score can be implemented as one of the criteria in determining the pass mark in OSCEs. Based on the evidence in this thesis, it is suggested that future CS assessments, specifically in OSCEs, should use a standardised instrument. Otherwise, a calibration process should be performed retrospectively to analyse the previous OSCE for quality assessment. The analysis of the OSCE's reliability using Generalizability Theory is also required. Hence, by understanding what types of error were produced in previous OSCEs, and subsequently Decision Study model, improvement could be made in an accurate and efficient manner.

6.7. Dissemination and Other Achievements During the Project

We commenced this research project in September 2012. In the course of the project, a number of outputs were achieved. To date (November 2015), two papers have been published, and another paper has been submitted as a direct result of the study. Several abstracts, posters and a paper were presented at various conferences. Below is a list of such papers, abstracts and conferences, in chronological order. The presenter is indicated by '(P)'.

- Abstract: The use of OSCEs to assess communication skills in undergraduate medical students: A systematic review of published literature. Winny Setyonugroho (P), Kieran M Kennedy, Thomas JB Kropmans. The Association for Medical Education in Europe (AMEE) 2013 conference. September 2013. Oral presentation.
- Abstract: The use of OSCEs to assess communication skills in undergraduate medical student: A systematic review of published literature. Winny Setyonugroho (P), Kieran M Kennedy, Thomas JB Kropmans. Galway Alliance of Medical Educators (GAME) meeting. December 2013. Oral presentation.
- Abstract: The use of OSCEs to assess communication skills in undergraduate medical student: A systematic review of published literature. Winny Setyonugroho (P), Kieran M Kennedy, Thomas JB Kropmans. Asia Pacific Medical Education Conference (APMEC) 2014. January 2014. Poster presentation.
- Abstract: The use of OSCEs to assess communication skills in undergraduate medical student: A systematic review of published literature. Winny Setyonugroho (P), Kieran M Kennedy, Thomas JB Kropmans. NUIG research day. Poster presentation.
- Abstract: Calibration of Communication Skills Items According To The Maas-Global In Objective Structured Clinical Examinations: A Pilot Study. Winny Setyonugroho (P), Kieran M Kennedy, Thomas JB Kropmans. NUIG research day. Poster presentation.
- Abstract (co-author): Spy glasses prove paper marking OSCE stations is faster than using tablets, the truth. Ayisha Hennely (P), Thomas JB Kropmans, Winny Setyonugroho. Summer research project. Poster presentation.
- Abstract: Calibration of Communication Skills Items According To The Maas-Global In Objective Structured Clinical Examination. Asia Pasific

Medical Education Conference (APMEC) 2015. Winny Setyonugroho (P), Thomas JB Kropmans, Kieran M Kennedy, Brian Stewart, Jan van Dalen. Asia Pasific Medical Education Conference (APMEC) 2015. February 2015. Poster presentation.

- Abstract: Fixing the rubber band: calibration of communication skills items in OSCE checklists. Winny Setyonugroho (P), Thomas JB Kropmans, Kieran M Kennedy, Brian Stewart, Jan van Dalen. Irish Network Medical Educators (INMED) 2015. Oral presentation.
- Abstract: Fixing the Rubber Band in Communication Skills Assessment of Interdepartmental OSCE Stations. Winny Setyonugroho (P), Thomas Kropmans, Ruth Murphy, Peter Hayes, Kieran M Kennedy. 5th International Nursing and Midwifery Conference 2015.
- Paper: Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review. Patient Educ Couns 2015 [13].
- Paper: Calibration of Communication Skills Items in OSCE Checklists according to the MAAS-Global. Patient Educ Couns 2015 [28].
- Paper (co-author): Back to the future: An online OSCE Management Information System for nursing OSCEs. Nurse Educ Today (2015) [29].
- Paper: Language Proficiency and Student Performance in English-speaking Country Medical Schools: A Mini Review. Winny Setyonugroho, Iman Permana (P). International Conference of Medical and Health Sciences (ICMHS) 2015. Oral presentation.
- Abstract: Calibration of Communication Skills Items According To The Maas-Global In Objective Structured Clinical Examination. Winny Setyonugroho (P), Thomas JB Kropmans, Kieran M Kennedy, Brian Stewart, Jan van Dalen. The Association for Medical Education in Europe (AMEE) 2015 conference. September 2015. Oral presentation.

- Paper (submitted): True Communication Skills Assessment in Interdepartmental OSCE Stations : Comparison of Multiple Instruments. Winny Setyonugroho, Thomas Kropmans, Ruth Murphy, Peter Hayes, Kieran M Kennedy.
- Paper (submitted): The effect of language on the assessment of communication skills using a standardized measurement instrument in OSCEs. Winny Setyonugroho, Maureen Kelly, Kieran M Kennedy, Eva Flynn, Rosemary Geoghegan, Ruth Murphy, Peter Hayes, Thomas Kropmans.

6.8. References

- [1] J. Cardinet, S. Johnson, G. Pini, *Applying Generalizability Theory Using Edug*, Taylor & Francis, 2012.
- [2] S.M. Downing, Reliability: on the reproducibility of assessment data, *Med. Educ.* 38 (2004) 1006–1012.
- [3] J.C.G. Jacobs, E. Denessen, C.T. Postma, The structure of medical competence and results of an OSCE, *Neth. J. Med.* 62 (2004) 397–403.
- [4] J. van Dalen, E. Kerkhofs, B.W. van Knippenberg-van den Berg, H.A. van den Hout, A.J.J.A. Scherpbier, C.P.M. van der Vleuten, Longitudinal and concentrated communication skills programmes: Two Dutch medical schools compared, (2002).
- [5] S. Scheffer, I. Muehlinghaus, A. Froehmel, H. Ortwein, Assessing students' communication skills: validation of a global rating, *Adv. Health Sci. Educ. Theory Pract.* 13 (2008) 583–592. doi:10.1007/s10459-007-9074-2.
- [6] H.M. Bosse, J.-H. Schultz, M. Nickel, T. Lutz, A. Möltner, J. Jünger, et al., The effect of using standardized patients or peer role play on ratings of undergraduate communication training: a randomized controlled trial, *Patient Educ. Couns.* 87 (2012) 300–306. doi:10.1016/j.pec.2011.10.007.
- [7] P.H. Harasym, W. Woloschuk, L. Cuning, Undesired variance due to examiner stringency/leniency effect in communication skill scores assessed in OSCEs, *Adv. Health Sci. Educ. Theory Pract.* 13 (2008) 617–632. doi:10.1007/s10459-007-9068-0.
- [8] J. Silverman, The Calgary-Cambridge guides: the “teenage years,” *Clin. Teach.* 4 (2007) 87–93.
- [9] J. Burt, G. Abel, N. Elmore, J. Campbell, M. Roland, J. Benson, et al., Assessing communication quality of consultations in primary care: initial reliability of the Global Consultation Rating Scale, based on the Calgary-Cambridge Guide to the Medical Interview, *BMJ Open.* 4 (2014) e004339.

- [10] H. Boon, M. Stewart, Patient-physician communication assessment instruments: 1986 to 1996 in review, *Patient Educ. Couns.* 35 (1998) 161–176. doi:10.1016/S0738-3991(98)00063-9.
- [11] R.S. Beck, R. Daughtridge, P.D. Sloane, Physician-patient communication in the primary care office: a systematic review., *J. Am. Board Fam. Pract.* 15 (2002) 25–38.
- [12] J. van Thiel, P. Ram, J. van Dalen, MAAS-global manual, Maastricht Maastricht Univ. (2000) 4–5.
- [13] W. Setyonugroho, K.M. Kennedy, T.J.B. Kropmans, Reliability and validity of OSCE checklists used to assess the communication skills of undergraduate medical students: A systematic review, *Patient Educ. Couns.* (2015). doi:10.1016/j.pec.2015.06.004.
- [14] Y. Salamonson, B. Everett, J. Koch, S. Andrew, P.M. Davidson, English-language acculturation predicts academic performance in nursing students who speak English as a second language, *Res. Nurs. Health.* 31 (2008) 86–94. doi:10.1002/nur.20224.
- [15] T. Harvey, C. Robinson, R. Frohman, Preparing culturally and linguistically diverse nursing students for clinical practice in the health care setting, *J. Nurs. Educ.* 52 (2013) 365–370. doi:10.3928/01484834-20130529-02.
- [16] T.T. Donnelly, E. McKiel, J.J. Hwang, Challenges and motivators influencing the academic performance of English as an additional language (EAL) nursing students: the perspectives of the students, *CJNR Can. J. Nurs. Res.* 41 (2009) 130–150.
- [17] C. Alptekin, G. Erçetin, The role of L1 and L2 working memory in literal and inferential comprehension in L2 reading, *J. Res. Read.* 33 (2010) 206–219. doi:10.1111/j.1467-9817.2009.01412.x.
- [18] T. Crawford, S. Candlin, A literature review of the language needs of nursing students who have English as a second/other language and the effectiveness of English language support programmes, *Nurse Educ. Pract.* 13 (2013) 181–185. doi:10.1016/j.nepr.2012.09.008.
- [19] L.A. Baig, C. Violato, R.A. Crutcher, Assessing clinical communication skills in physicians: are the skills context specific or generalizable, *BMC Med. Educ.* 9 (2009) 22. doi:10.1186/1472-6920-9-22.
- [20] A. Chur-Hansen, Language background, proficiency in English, and selection for language development, *Med. Educ.* 31 (1997) 312–319.
- [21] W. Setyonugroho, I. Permana, Language Proficiency and Student Performance in English-speaking speaking Country Medical Schools: A Mini Review, in: *Int. Conf. Med. Health Sci. 2015*, Universitas Muhammadiyah Yogyakarta, Indonesia, 2015. <http://iiste.org/Journals/index.php/JHMN/article/view/25374>.
- [22] C. Mann, B.J. Canny, D.H. Reser, R. Rajan, Poorer verbal working memory for a second language selectively impacts academic achievement in university medical students, *PeerJ.* 1 (2013) e22. doi:10.7717/peerj.22.

- [23] J. Zhang, L. Xie, Y. Li, M. Chatterjee, N. Ding, How Noise and Language Proficiency Influence Speech Recognition by Individual Non-Native Listeners, *PLoS ONE*. 9 (2014) e113386. doi:10.1371/journal.pone.0113386.
- [24] H.S. Brand, M. Schoonheim-Klein, Is the OSCE more stressful? Examination anxiety and its consequences in different assessment methods in dental education, *Eur. J. Dent. Educ.* 13 (2009) 147–153. doi:10.1111/j.1600-0579.2008.00554.x.
- [25] R. Bloch, G. Norman, Generalizability theory for the perplexed: A practical introduction and guide: AMEE Guide No. 68, *Med. Teach.* 34 (2012) 960–992. doi:10.3109/0142159X.2012.703791.
- [26] M.T. Brannick, H.T. Erol-Korkmaz, M. Prewett, A systematic review of the reliability of objective structured clinical examination scores, *Med. Educ.* 45 (2011) 1181–1189. doi:10.1111/j.1365-2923.2011.04075.x.
- [27] C.L. Kimberlin, A.G. Winterstein, Validity and reliability of measurement instruments used in research, *Am. J. Health. Syst. Pharm.* 65 (2008) 2276–2284. doi:10.2146/ajhp070364.
- [28] W. Setyonugroho, T. Kropmans, K.M. Kennedy, B. Stewart, J. van Dalen, Calibration of Communication Skills Items in OSCE Checklists according to the MAAS-Global., *Patient Educ. Couns.* (n.d.). doi:10.1016/j.pec.2015.08.001.
- [29] P. Meskell, E. Burke, T.J.B. Kropmans, E. Byrne, W. Setyonugroho, K.M. Kennedy, Back to the future: An online OSCE Management Information System for nursing OSCEs, *Nurse Educ. Today*. (2015). doi:10.1016/j.nedt.2015.06.010.

Appendix 1. MAAS-Global check-list calibration manual

Maas-Global Check-lists Calibration Manual

Introduction

Heterogeneity in measurement instruments used to assess Communication Skills (CS) in OSCEs limits the comparability of student performance between examinations (Setyonugroho et al, 2013). Comparison of results obtained from different instruments is impossible because each instrument measures different phenomena.

Our previous research found that Maas-Global has been used by several institutions as an international reference standard for CS.

In other studies, the Maas-Global has been reviewed as good instruments for measuring CS in undergraduate medical education settings.

The aim of this project is to calibrate existing CS stations of the School of Medicine of the NUI Galway. Your assistance as clinical tutor in this grading process is kindly appreciated.

Calibration

- To calibrate Communication Skills content of CS assessment forms on the international standard Maas-Global.
- Using the Maas-Global Calibration Checklists (MGCC) to rate how close items in station's checklists fits the golden standard.
- MGCC is the Maas-Global checklists which adapted to calibrate CS items by adding numbering (alphabet) in sub-items.

Maas-Global Calibration Checklist (MGCC)

- Consist of 3 sections and 17 items
7 items in section 1 (Communication skills for each separate phase of a consultation), 6 items in section 2 (General communication skills – how is done), and 4 items in section 3 (Medical Aspect – what is done)
- In section 1 and 2, each item contain 3 to 4 sub-items.
- Items 1-13 concern communication skills.
Items 14 – 17 for rating the medical content of consultation. It is the student's competence that is being rated both qualitatively and quantitatively.

SECTION 1: COMMUNICATION SKILLS FOR EACH SEPARATE PHASE	
1 INTRODUCTION	
a. giving the patient room to tell his story	5 DIAGNOSIS
b. general orientation on the reason for visit	a. naming findings and diagnosis/hypothesis
c. asking about other reasons for visit	b. naming causes or the relation between findings and diagnosis
2 FOLLOW-UP CONSULTATION	c. naming prognosis or expected course
a. naming previous complaints, requests for help and management plan	d. asking for patient's response

How to calibrate

1. Mark the items of the NUIG check-lists according to the Maas-Global. Definition of MGCC is taken from the Maas-Global Manual.
2. Read the MGCC manual (this take approximately 1.5 hours)
3. Write the item number.
eg: 3 , 7, 11, etc If an item is purely about a non communication skills, it should be mark as N/A

Example :

Skills Communication (14 marks)	Performed adequately and completely	Attempted but inadequate or incomplete	Not attempted or grossly incorrect
Introduces himself to patient (2 marks)	3		
Explains the procedure (4 marks)	4		
Obtains verbal consent (4 marks)	4		
Offers a chaperone (4 marks)	3		
Skills Clinical (38 Marks)	Performed adequately and completely	Attempted but inadequate or incomplete	Not attempted or grossly incorrect
Indicates that he would ask the patient to undress from the waist down (2 marks)	4		
Indicates that he/she would turn around and offer privacy while patient is undressing (2 marks)	4		
Position of patient for examination Indicates that he would ask the patient to lie in the left lateral position i.e. to lie on the left side, knees drawn up towards the chest and lumbar region on the edge of the bed.	4		
Wears gloves	N/A		

DEFINITIONS OF CONCEPTS

Communication skills

Communication skills in this manual concern behaviour of the doctor that is conducive to effective communication. They are mostly verbal skills but also some non-verbal skills. Doctor and patient communicate effectively if both seek to bring their mutual goals into line with each other and if both are aware of the meaning of the information exchanged.

Comments:

The emphasis is on the form of communication. The doctor's skills in this regard are: asking questions, summarizations, reflecting feelings, ordering, structuring, exploring requests for help, information sharing, and involving the patient in the matter under consideration.

Asking (additional) questions (items 1, 2, 5, 6, 7, 8, 9, 10, 11, 14)

Asking (additional) questions refers to the doctor using inviting phrases to encourage the patient to tell something more about a topic.

Comments:

Linguistically, the grammatical question (such as "how much does it bother you?") is the best form to be used in asking for information. It is, however, rarely used in everyday conversation. Most people only need a declarative sentence ("It is a problem?"), a paraphrase ("Uncomfortable?") or a literal repetition ("Bothersome") as an invitation to tell more. Whether the doctor intended a remark as a question can only be confirmed by asking the doctor, since the context of a consultation rarely provides enough information to determine this accurately.

Furthermore, one patient may regard a remark as an invitation, whereas another patient may interpret it as a sign that the doctor has understood what was said. In light of these considerations observers should bear in mind that the act of "asking" can be represented as a proper question, a declarative sentence or a literal repetition.

No distinction is made in the MAAS-Global between open-ended, closed-ended and leading questions. Whether the patient experiences an open question as an invitation to talk freely, depends largely on the context in which the question is asked. Indeed, in some cases an open question may even be experienced as threatening. A closed question, on the other hand, may invite the patient to tell more. With leading questions the context also partly determines to what extent the patient feels inhibited by the suggestion.

Exploring requests for help (items 3 and 9)

Exploring of requests for help refers to exploration (item 9) and naming (item 3) of the following key questions:

1. What change in the present situation is expected by the patient.
2. What are the patient's wishes and expectations about how this change can be brought about and the doctor's role in this respect.

The result of 1 and 2 is called request or requests for help.

Comments:

Key question 1: The patient experiences the present situation as one that is undesirable, and unacceptable. Usually, it involves discomfort or pain, physical and/or mental. The patient wants to change this situation and does or does not have a very clear picture of how to bring this about. In many cases patients have not given much thought to the desired situation and the doctor has to help the patient explore what he wants by asking clarifying questions.

Key question 2: The patient will have wishes or expectations regarding the course of action to be taken to change the present situation into the desired one. These wishes or expectations may be clear before the patient sees the doctor, they may also become clear through questions asked by the doctor, or they may gradually become clear over the course of their encounters. Wishes and expectations may concern both what the doctor will do during the diagnostic phase (history-taking and physical examination) and the plan proposed during the management phase (wait and see, treatment, referral, etc.) Wishes and expectations can also be negative, for example the patient's expectation that the doctor will write another prescription he does not want, or that he will be referred to a specialist he does not want to see as happened on the previous visit.

The following are some of the issues that may be related to one or the other of the above key questions. However, their importance will vary depending on the case concerned:

1. What considerations prompted the patient to seek help at this particular time.
2. What are the patient's feelings concerning the complaint or problem.
3. What are the patient's suspicions or assumptions with regard to the cause of the complaint or problem.
4. What has the patient himself done to manage the undesirable condition.
5. What impact have significant others (partner, family, friends) or important living situations (work, hobbies, sports) had on all of the above questions?

When the doctor introduces the above subjects himself, this does not automatically constitute exploration of requests for help. The patient's reaction, how the doctor pursues the subject and the result determine whether it concerns a request for help. For example, when the doctor asks: "What does your husband think about it" the patient may answer: "Well, all he ever thinks about are his pigeons", or "he mentioned cancer and now I am very worried". The first answer is unlikely to be related to the patient's request for help, the second answer can lead to the request "I need reassurance".

The first of the above issues (what considerations prompted the patient to seek help) is relevant to the request for help in so many cases that it is a subitem of the item on requests for help.

Exploring requests for help is not synonymous with attention to psychosomatic aspects or psychosocial background, although these may be considered in connection with the request for help. For example: although the fact that his partner gives the patient headaches is a psychosocial factor, this does not necessarily mean that it has to be discussed in the context of requests for help. In this patient's case exploring requests for help might reveal that he wants the doctor to prescribe a painkiller for his headache and that his psychosocial problem is not a request for help. If the doctor thinks it necessary, the psychosocial background can be dealt with during the (psychosocial) history and should be rated there (item 14).

There are situations where exploring requests for help is less appropriate and the doctor may limit this or not use it at all, such as in an emergency, a telephone consultation, when there are major language problems or large cultural differences.

The consequences of this for measurement are addressed in the remarks on selection of consultations in the sections on validity and reliability in the introduction.

Frame of reference (item 8)

A frame of reference is a set of customs, a pattern of standards, one's perspective on reality, coloured glasses through which one looks at the world.

Comments:

The patient's frame of reference is the whole of his notions and perceptions concerning the complaint or problem. It is also referred to as the patient's perceptions of reality or the patient's perspective. It is always determined by highly personal factors and can only be discovered by responding to remarks (such as "one hears all kinds of things", "I am rather worried", "do you think that something else is wrong?") or by the doctor directly asking questions

("are you worried?", "what do you think is wrong?").

The doctor's frame of reference is the whole of his notions and perceptions pertaining to his work as a professional. Apart from individual characteristics, the doctor's frame of reference generally includes the medical model of history, physical examination, diagnosis and management.

Reflection of feelings (item 9)

Reflection of feelings means that the doctor gives a verbal rendering of the feelings expressed by the patient during the consultation either in words or nonverbally. The doctor's reflection must:

- appropriately reflect the nature of the feelings
- accurately reflect the intensity of the feelings

Comments:

Example: A patient has just told the doctor that he is terrified of having an operation and the doctor responds by saying: "I understand that you feel some concern". Although the doctor does reflect the patient's feelings, he does not accurately label the feeling (concern rather than terror) or reflect its intensity (some versus a large amount). A better response would have been: "You are really very anxious".

Doctors often say something like "I can understand that" to show that they sympathize. The above definition shows that such a remark is not a reflection of feelings. At best it shows empathy, at worst it is a cliché. It becomes a reflection of feelings if subsequently the feeling is named, such as "I can understand that having this operation is really very frightening".

Summarizations (item 11)

Summarization means that the doctor rephrases the main topics introduced by the patient in the preceding part of the consultation. The main purpose is to check whether the doctor has accurately understood the patient's intentions.

Comments: A good summarization should meet the following requirements:

- it should accurately reflect the content of what the patient said
- it should be concise: a brief account of what the patient has said
- it should be a re-phrasing of the account rather than a literal repetition
- it should seek verification of the summarization by directly asking the patient for verification, by using a questioning tone of voice, or by following the summarization with a pause that invites the patient's response.

SECTION 1: COMMUNICATION SKILLS FOR EACH SEPARATE PHASE

Item 1: INTRODUCTION

- giving the patient room to tell his story
- general orientation on the reason for visit
- asking about other reasons for visit

In the initial phase of the consultation the student orientates himself with regard to the reason for the visit by giving the patient room to talk about his complaints, problems or questions in his own words and, if necessary, by asking general questions to encourage the patient. General questions include questions about how long the patient has had the problem or complaint, how serious it is and what it means to the patient. The opening question is not rated. The student explores whether there are any other reasons for the patient's visit. In rating this aspect the timing of this question is crucial: before starting detailed history-taking.

Comments:

"Reason for visit" includes anything that is initially brought up by the patient, such as complaints, problems or questions. Occasionally, it may be difficult to distinguish between the orientation on the reason for the visit and questions pertaining to the history. The main distinction is that orientation is concerned with the main points of the history, not the details. Examples of orientating questions are: What is the problem? How long have you had this? How much does it bother you? These questions help the student to orientate about the patient's reasons for visit. History-taking involves specific questions from the student's perspective. The more the opening questions resemble history related questions the lower the rating on this item. Questions pertaining to the history are rated in item 14 (history-taking).

Although questions about other reasons for the visit should not be among the opening questions, they should be asked quite early in the consultation. These questions are included in the introduction for reasons of organisation. It is important to ask about other reasons for the visit to ensure that these reasons are not overlooked, help the patient complete his story and enables the student to plan the consultation. The opening question ("Well, tell me", "What can I do for you") is not rated because this question is almost always asked and therefore does not contribute to differences in ratings between students.

Item 2: FOLLOW-UP CONSULTATION

- naming previous complaints, requests for help and management plan
- asking about adherence to management plan
- asking about the course of the complaint

In a follow-up consultation the student makes the

connection with the previous consultation by naming the previous complaints, requests for help and arrangements made.

The student also finds out whether the patient has complied with the agreed management plan. The student also asks about the course of the complaint and the effect of the treatment or management strategy.

Comments:

A follow-up consultation is a sequel to a previous consultation with the same student about the same subject within the same illness episode. If one of this aspects is missing n.a. (not applicable) should be circled. Unless the context clearly shows otherwise, it should be assumed that the same student is involved. When the student summarizes issues from a previous consultation, for instance by reading from the record, the issues mentioned are rated here. The summarization as such is not rated, because summarization applies to issues of the present consultation that the patient has brought up (see definition in chapter 3).

Item 3: REQUEST FOR HELP

- naming requests for help, wishes or expectations
- naming reasons that prompted the patient to come now
- completing exploring request for help

The student names the patient's requests for help, wishes or expectations.

In addition the student names the reason the patient states why he came for the visit.

The student completes the request for help by checking whether all patient's questions, wishes or expectations have been addressed.

Comments:

For more details see the definition of request for help under "definitions of concepts".

This item pertains to the content of request for help, i.e. to what extent the student demonstrates that he has fully heard and acknowledged what the patient wants to say. When the student limits the request for help to asking questions and fails to rephrase the patient's responses, the rating is "0" on the first two aspects of this item.

The questions themselves and their quality are rated in item 8 (exploration).

"Naming requests for help" and "reasons that prompted the patient to come now" does not apply to suggestions made by the student that were not first expressed by the patient. For instance, the following would not result in a rating: student: "You want this symptom cleared up?", patient: "Well, what I really want to

know is whether it is serious". The important thing is that the student rephrases a topic that the patient has brought up. This type of question that includes a suggestion by the student can be rated as a suggestive way of exploring in item 8 (exploration).

Completion of request for help and the reason for presenting now can be inferred from the patient's affirmative response to a question like "So the main thing was ... and you expect me to ... Is that really everything you want?" This criterion has been added because students, if they pay any attention to request for help at all, tend to be easily satisfied when they clarify one request for help and do not ask for any further requests.

Exploration of request for help can occur or continue during any phases of the consultation. For instance, the student may ask the patient about his expectations regarding the management plan. It follows that this item can be rated during any phase of the consultation provided the student actually names requests for help, wishes or expectations. Naming the request for help is rated in this item, exploration should be rated in item 8 (exploration).

Item 4: PHYSICAL EXAMINATION

- a) instructions to the patient
- b) explanation of what is being done
- c) treating the patient with care and respect

The student tells the patient before he performs the physical examination where it will take place, which parts of the body should be uncovered and what the patient should do (lie, sit, etc.).

The student explains what the examination entails and explains his further actions during the examination if necessary.

The student treats the patient with care and respect. He anticipates the patient's reactions to the examination, e.g. pain, and addresses them. When no physical examination is performed, either indicated or not, "n.a." should be circled.

When, for any reason, no physical examination is performed, n.a. should be circled.

Comments:

Observers should not rate the medical aspects of the physical examination here, since this is rated in item 15. The announcement that the examination will be performed is rated in item 12.

Explanation of what is being done is limited to explaining what the student is doing not why it is being done. It does not include giving reasons or arguments why the examination is performed. With examinations that are more complicated, take more time or are more invasive the student may explain his further actions during the examination. With a limited examination, such as auscultation or measuring blood pressure it is enough to explain the procedure before

the examination is performed.

Item 5: DIAGNOSIS

- a) naming findings and diagnosis/hypothesis
- b) naming causes or the relation between findings and diagnosis
- c) naming prognosis or expected course
- d) asking for the patient's response

The student names the main findings from the history and physical examination, followed by a diagnosis or working hypothesis.

In addition the student tells about the causes of the complaint or disorder, or the connection between findings and diagnosis.

The student gives a concrete indication of the seriousness, the expected duration of the complaint and the course, with or without treatment.

Finally, the student asks the patient to give his reaction to the findings, diagnosis, prognosis etc.

Comments:

The diagnosis may also involve negative findings, such as "I cannot find anything unusual". A (preliminary) diagnosis is also rated when the student says that he is not able to draw any definite conclusions or when the diagnosis is formulated negatively, such as "It is definitely not a hernia".

Naming causes or the relation between findings and causes is rated regardless of the way in which the student does so.

When a student names a finding, diagnosis, cause, etc. in response to a question by the patient, the item is rated, unless it is evident that the student would not have named it without being prompted. An example of the latter would be when the patient asks on leaving: "How long will it take?"

When "asking for the patient's response", one question is enough for the behavior to be rated. The quality of the question and further exploration is rated in item 8 (exploration).

The content of the "diagnosis" phase should be disregarded. **Content is rated in item 16.**

Item 6: MANAGEMENT

- a) shared decision making, discussing alternatives, risks and benefits
- b) discussing feasibility and adherence
- c) determining who will do what and when
- d) asking for patient's response

The student discusses the management strategy by letting the patient have his say by asking the patient's opinion or by making an inviting pause. The risks and benefits of the proposed management strategy are also discussed. Depending on the nature of the complaint the student may need to discuss alternatives or indicate that there are no alternatives. The risks and benefits of

the proposed management strategy and any alternative strategies are also discussed.

The student talks about the feasibility of the proposed strategy taking into account the patient's possibilities and the student verifies if and to what extent the patient will adhere to the proposed management strategy.

The student makes concrete arrangements about further medical actions (who, what, when).

Finally, the student asks about the patient's reactions to the proposed course of action and arrangements.

Comments:

The patient's reaction as such is irrelevant. It is not important whether the patient reacts to the explanation of the risks and benefits or does not choose between alternatives offered by the student or leaves the decision to the student. The item is concerned with the invitation by the student, not with the patient's reaction. As with all other items the observer is only concerned with the student's behaviours in carrying out each of the subitems.

When the student carries out a subitem in response to a patient question, it is rated, unless it is evident that the student would definitely not have mentioned it otherwise, for instance when the patient asks on leaving: "Do I have to come again?"

In eliciting the patient's response to the management plan one question suffices. The quality of the question and any further explorations are rated in item 8 (exploration).

The medical content of the "management" phase

should be disregarded. It is rated in item 17.

Item 7: EVALUATION OF CONSULTATION

- a) general question
- b) responding to request for help
- c) perspective for the time being

At the end of the consultation the student asks a general question about what the patient thinks or feels at this moment. The question need not concern any specific aspect of the consultation.

At the end of the consultation the student checks whether the patient's requests for help have been adequately addressed.

The student checks whether the patient has been offered perspective for the time being.

Comments:

Evaluation of consultation can be rated on the basis of a general question that the student asks at the end of the consultation ("All right?", "Do you agree?", "Are you satisfied?") even though the student may not have intended this as an evaluation. Such questions often refer to what was discussed last, generally management and the arrangements made. In those cases the question is rated either in item 6 (management) or here.

Evaluation of the student's response to the request for help will depend on whether the student explicitly refers to the request for help. This may mean that the student refers to "your questions" or, preferably, actually names the requests for help.

Section 2: GENERAL COMMUNICATION SKILLS

Item 8: EXPLORATION

- a) exploring requests for help, wishes or expectations
- b) exploring patient's response to information given
- c) within patient's frame of reference
- d) responding to nonverbal behaviour and cues

The student explores the patient's requests for help, wishes or expectations by asking questions. This should be done in an inviting manner.

The student explores the patient's reaction to the information given. This applies in particular to the phases "diagnosis" and "management"

Exploration takes place within the patient's frame of reference.

While exploring the student responds to nonverbal behaviour and cues.

Comments:

This item measures the quality of the questions asked by the student to clarify the patient's perceptions of his complaints. These perceptions play an important part in patients' requests for help, wishes and expectations, and also in the way patients react to information given by the student. In **item 3** (request for help) only naming the requests for help, wishes and expectations is rated. Items **5** (diagnosis) and **6** (management) rate only whether the student asks the patient's response, **not how this is done**.

A prerequisite for exploration is that the student creates an inviting, open and safe climate for the patient. If he succeeds in doing so, open-ended questions are the best approach to exploration. However, sometimes the student may find he should ask only closed-ended questions. Thus it is not important what type of question (open or closed) is used, but rather whether the questions and the student's attitude are inviting. Since this item concerns the exploration of the patient's perceptions, so the student should ask

questions within the patient's frame of reference. Asking this type of question is by no means easy and the student may tend to ask questions from his own perspective. This is not prohibited, but it does not contribute to the rating. The student who best succeeds in keeping his questions within the patient's frame of reference, will obtain the highest ratings on this item. Exploration can be required in any phase of the consultation, but it is most appropriate during request for help, diagnosis and management. During these phases the observer should be especially on the alert for rating exploration as well as the items concerned (3, 5 and 6).

Exploration should be relevant within the context of the complaint or the student should explain the relevance of the question to the complaint. For example a student may explore personal or psychosocial conditions within the patient's frame of reference, even though the patient did not present these as complaints nor has any idea why the student explores these issues.

For further comments see also the definitions of exploration and frame of reference.

Item 9: EMOTIONS

- a) asking about/ exploring feelings
- b) reflecting feelings (including nature and intensity)
- c) sufficiently throughout the entire consultation

The student asks about the patient's feelings or he asks questions when the patient shows emotions.

The student reflects the feelings that the patient shows and expresses appropriately, with respect to both their nature and intensity.

The student pays attention to the feelings throughout the consultation by asking questions and reflecting feelings sufficiently and with an appropriate balance of time, i.e. not too much and not too little.

Comments:

See for further explanation the definition of reflection of feelings under "definitions of concepts".

This item does not measure whether the interaction can be qualified as "cold" or "warm", or whether the patient is emotional or the student empathetic. The student's behaviours in this regard are often nonverbal and are rated in the item empathy (item 13). In rating this item "emotions" observers should rate how well the student responds verbally to patient emotions. Feelings and reflection of feelings concern the patient's feelings and emotional responses related to the complaint. Observers should not include the feelings of pain or discomfort associated with the complaint. These feelings are usually addressed during history-taking. It goes without saying that the item is not about the student's feelings.

Students often reflect patient emotions in

summarizations, which makes it difficult to observers to identify these as reflections of feelings. Nevertheless they should be rated separately from summarizations as reflections of feelings in this item.

Item 10: INFORMATION GIVING

- a) announcing, categorizing
- b) in small quantities, concrete explanations
- c) understandable language
- d) asking whether the patient understands

The student announces to the patient that he is going to give information about a subject and explains which categories will be dealt with.

The information is given in small quantities and the student explains details concretely.

The student uses language that is easy to understand for this particular patient.

The student checks whether the patient has understood the information by asking questions.

Comments:

Example of an announcement: "I will tell you something about what I have found".

Example of categorizing: "First I am going to tell you what I have found, then I will explain what I think is going on and finally I will explain what I think should be done about it. Now, first of all....."

"Small quantities" implies that the student does not give too much information at once. He can do this by pausing between pieces of information to give the patient the chance to absorb the information or to ask for clarification.

Students rarely announce that they are going to give information or state the categories of information they are going to give. They also rarely ask whether the patient has understood the information ("What do you think?", "Do you understand what I am saying?").

Observers should be aware of this, since it affects the rating.

Information giving only applies to the phases "diagnosis" and "management".

During these phases the observer should be especially alert and rate also the subject of the information (items 5 and 6) and the medical content (items 16 and 17), as well as this item 10 itself.

Item 11: SUMMARIZATIONS

- a) content is correct, complete
- b) concise, rephrased
- c) checking
- d) sufficiently throughout the entire consultation

The student demonstrates throughout the consultation that he has heard what the patient has to say through sufficient and well-balanced summarizations, phrases concisely, in his own words, contentwise correct, and he offers the patient room to respond (pause,

questioning intonation, asking question).

Comments:

See for further explanation the definition of summarization under definitions of concepts. The rating "excellent" can only be given when summarizations are integrated in the whole consultation in a well-balanced manner and when they are of good quality.

Summarizations of a previous consultation that occur in a follow-up consultation, should not be rated here. A summarization refers exclusively to what has been discussed in the present consultation (see definition).

Item 12: STRUCTURING

- a) logical sequence of phases
- b) balanced division of time
- c) announcing (history-taking, examination, other phases)

The student gives guidance to the consultation by ordering phases in a logical way, consecutively: introduction, follow-up consultation, request for help, history, physical examination, diagnosis, management and evaluation.

The student also divides his time between the phases used in a well-balanced way and, if necessary, intervenes to cut the story of a very talkative patient short. The student brings structure to the consultation by announcing the phases used.

Comments:

When the student leaves out a particular phase of the consultation, this should not negatively influence the rating, because "balanced division of time" applies only to the phases used.

With a very talkative patient it may be necessary and effective to interrupt the patient. If the student is too lenient and as a result of this runs out of time, time will no longer be distributed in a well-balanced manner over the different phases of the consultation. The student may do a part of the history during the physical examination provided he does so in an orderly manner. The "management" phase must follow the "diagnosis" phase.

Item 13: EMPATHY

- a) concerned, inviting and sincerely empathetic in intonation, gesture and eye contact
- b) expressing empathy in brief verbal responses

The student's attitude is inviting and shows his concern for the patient. Also he is sincere in showing empathy. This attitude is reflected in gestures, eye contact and tone of voice.

The student expresses empathy in brief verbal responses.

Comments:

Empathy refers to concern and sympathy. It comprises verbal and nonverbal aspects.

Nonverbal expressions of empathy are observed when the student displays a clearly patient-centred attitude and speaks in a tone of voice that shows real empathy and is supported by appropriate gestures and eye contact.

All nonverbal expressions of empathy are rated in this item.

Verbal expressions of empathy are seen in behaviour that is partly rated in other items, for instance treating the patient with care and respect (item 4), exploring requests for help, wishes and expectations (item 8) and asking about / exploring feelings, and reflecting feelings (item 9). Verbal responses that are to be rated in this item 13 are those when the student briefly repeats what the patient has said to indicate that he is listening or short responses like "uh huh" or "mm", to show that he is listening or to encourage the patient to go on. Short expressions like "oh really" or "that's awful", which are clearly intended to show sympathy are also rated in this item.

Other evidence of empathy:

- not interrupting the patient without good reason
- conducting the conversation in a quiet environment by avoiding unnecessary interruptions (telephone conversations, people coming and going)
- avoiding awkward silences
- not starting a lengthy conversation when the patient is undressed or is undressing.

Empathy should be evident from the student's behaviour. Empathy cannot be inferred from the fact that the patient appears to feel at ease.

In rating empathy a variety of behaviours should be considered: verbalisations, intonation (calm, inviting) and posture (directed toward the patient, eye contact while speaking, the student's gestures in greeting the patient and when the patient is leaving). It will be clear that the observer should not only listen, but also watch carefully.

In observing a videotaped consultation eye contact could be difficult to judge. In this case it should suffice that the student's body and head are turned toward the patient and that the student is not writing, typing on the computer etc. while talking to the patient.

Leniency in management aspects is not a sign of empathy and its consequences should be rated in item 17 (management).

Section 3: MEDICAL ASPECTS

The items in this section (item 14-17) are intended for rating the medical content of the consultation. Items 1-13 concern communication skills, items 14-17 are related to what the student says and does as a medical professional. It is the student's medical competence that is being rated, both quantitatively and qualitatively. Does the student ask the right questions and is the number of questions adequate? Is the physical examination appropriate? Does the student explain his findings to the patient adequately and accurately? Is the management plan in line with professional guidelines?

The comments in this chapter were written with Dutch GPs in mind, since they are currently the main group of users. If the need arises, the comments have to be adapted to other groups of users.

When the observer's professional organisation or medical society has published guidelines for specific diseases or complaints, consultations involving these disorders should be rated in accordance with these guidelines. Consultations involving disorders for which no guidelines have been published should be rated in accordance with the prevailing professional standard. In these cases the rating is more difficult to assess.

Unlike the items on communication skills the items on medical aspects lack subitems. This is due to the uniqueness of a case, i.e. what is obligatory in one case, like asking a specific question during history taking, the physical examination or the management strategy, may be completely irrelevant in another one. The subitems in the MAAS-Global are applicable to all cases. Another reason why no subitems are included is that some concepts in medicine are not firmly grounded in evidence nor clearly defined. This applies for instance to the "psychosocial history", "asking questions about psychosocial aspects", "attention to psychosocial consequences". Furthermore these aspects are not relevant for all cases, although opinions about this may differ. For these reasons items on medical content have no subitems and are presented only with a list of aspects that may be rated.

When consultations are rated with a view to feedback and educational purposes, it is advisable to note the medical aspects that have strongly affected the rating under other feedback. For instance in the history: "the distribution of attention over somatic, psychological and social aspects was well-balanced" or regarding management: "drug therapy not indicated according to the guideline".

Item 14: HISTORY-TAKING

This item can be used to rate somatic history and psychosocial history, if applicable.

Rate according to professional guidelines if they are available. Otherwise rate to the best of your ability.

Comments:

If a psychosocial history is appropriate, but not obtained, the rating should be lower, regardless of the quality of the somatic history.

Item 15: PHYSICAL EXAMINATION

This item can be used to rate if applicable:

- physical examination by the student
- additional tests done by the student during the consultation

Rate according to professional guidelines if they are available. Otherwise rate to the best of your ability.

Comments:

Physical examination consists of the examination and additional investigations carried out during the consultation. Additional investigations that are planned after the consultation are rated under "management" (item 17).

Physical examination that is not recommended in the guidelines is considered superfluous and should result in a lower rating.

If data obtained in the history or in previous consultations indicate that a physical examination is not necessary, raters should circle "n.a."

Item 16: DIAGNOSIS

This item can be used to rate diagnosis or working hypothesis.

Rate according to professional guidelines if they are available. Otherwise rate to the best of your ability.

Comments:

The observer rates the medical quality of the "diagnosis" phase using the information that the student gives to the patient. This concerns the phase when the student makes his diagnosis. The student decides which diagnosis or working hypothesis to use on the basis of the findings from the history and the physical examination, or he decides that he does not know. All this takes place inside the student's head and it is only shown to the observer and the patient when the student tells his findings, considerations, diagnosis, causes, prognosis and expected course of disease. This item is concerned with the medical content of the diagnosis.

Item 17: MANAGEMENT

For this item observers should rate the following aspects if applicable:

- wait and see
- education
- treatment
- medication
- additional tests
- referral

Rate according to professional guidelines if they are available. Otherwise rate to the best of your ability.

Comments:

Medication and other treatment strategies fall under "management". When appropriate, education is also a part of "management".

Any referrals and additional tests are included in the rating. If referral is indicated (by consulting guidelines?) this will lead to a higher rating. An inappropriate referral, i.e. referring the patient when this is not indicated, leads to a lower rating.

The patient's contribution may affect the choice of management strategy. The observer should take this into account when the student deviates from the management proposed in guidelines. If the student allows interpersonal factors to interfere with his adherence to consensus in management decisions, such as in cases where the student tries to avoid a conflict with the patient, this should have a negative effect on the rating.

Appendix 2. Ethical approval



24th August 2015

Dear Mr Setyonugroho

Re. Ethics Application: Untidy the Rubber Band: Comparison of Communication Skills Outcome in OSCEs

I write to you regarding the above proposal which was submitted for Ethical review. Having reviewed your response to my letter, I am pleased to inform you that your proposal has been granted **APPROVAL**.

All NUI Galway Research Ethic Committee approval is given subject to the Principal Investigator submitting annual and final statements of compliance. The first statement is due on or before 30th November 2015. Please see section 7 of the REC's Standard Operating Procedures for further details which also includes other instances where you are required to report to the REC.

Yours Sincerely

Allyn Fives

Chair, Research Ethics Committee

OE Gaillimh,
Bóthar na hOllscoile,
www.nuigalway.ie/research-support-services
Gaillimh, Éire

NUI Galway,
University Road
Galway, Ireland

T: +353 91 495969
F: +353 91 494951
E: rss@nuigalway.ie