Patient Education and Counseling 99 (2016) 139–146

Contents lists available at ScienceDirect

Patient Education and Counseling

journal homepage: www.elsevier.com/locate/pateducou

Healthcare education

Calibration of communication skills items in OSCE checklists according to the MAAS-Global

Winny Setyonugroho[a], Thomas Kropmans[b,*], Kieran M Kennedy[b], Brian Stewart[b], Jan van Dalen[c]

[a] Faculty of Medicine and Health Sciences of the Universitas Muhammadiyah Yogyakarta, Indonesia
[b] School of Medicine, College or Medicine, Nursing & Health Sciences, National University of Ireland Galway, Ireland
[c] Skills laboratory, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands

ARTICLE INFO

Article history:
Received 20 October 2014
Received in revised form 24 July 2015
Accepted 2 August 2015

Keywords:
Validity
Reliability
Communication skills
OSCE
Objective structured clinical examination
MAAS-Global

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 behaviours.
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.

1. 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 standardisation 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

* Corresponding author at: School of Medicine, Comerford Building Room 204, Clinical Science Institute, National University of Ireland Galway, Galway, Ireland. Tel.: +353 91 494340; fax: +353 91 495512.
E-mail address: thomas.kropmans@nuigalway.ie (T. Kropmans).

http://dx.doi.org/10.1016/j.pec.2015.08.001
0738-3991/

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.

2. Methods

2.1. 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.

2.2. 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 min with 1 min in between stations and 1 min 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 analysed. In total, 250 assessment forms used in 27 OSCE circuits (Table 1) were analysed. Further details of contributions from each discipline are presented in Table 1.

2.3. 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, summarisations, 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 A).

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").

2.4. 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 examiners(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 generalisability 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 Generalisability Theory study, analysis will more appropriately show the contribution of each of the potential sources of error variance to the total error [22]. The Generalisability Theory (G-theory) analysisis complementary to the classical psychometric theory and consists of a Generalisability-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.

Table 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
	2012–2013	2	10	229
Discipline of General Practice	2009–2010	1	8	117
	2010–2011	2	10	349
	2011–2012	2	10	231
	2012–2013	2	10	202
Discipline of Psychiatry	2009–2010	1	4	75
	2010–2011	2	8	170
	2011–2012	2	10	250
	2012–2013	2	10	195
Total				2577

2.5. 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 (Fig. 1). The upper portion of Fig. 1 provides an example of how each individual rater matched OSCE checklist items with the items of the MAAS-Global.

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".

2.6. 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 (Fig. 1). The lower portion of Fig. 1 provides an example of binary translation of the raters assessment forms.

The data was analysed as a 4-facet Generalisability 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 analyse the reliability of the calibration process according [18]. G-Kappa is the term used in generalisability theory to analyse binary data [23]. The accepted G-Kappa value for precision of measurement is 0.80 [18].

3. Results

3.1. 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.

3.2. Reliability analysis

Table 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 Generalised Kappa (G-Kappa) of the station checklists calibration, from all disciplines, was 0.8. D-study analysis

Skills Communication (14 marks)	Performed adequately and completely	Attempted but inadequate or incomplete	Not attempted or grossly incorrect
Introduces himself to patient (2 marks)	1		
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 in 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		

No.	Maas-Global Item																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Fig. 1. Illustration on how raters calibrate station's checklists and then transferred into spreadsheet.

demonstrated a decrease in G-Kappa from 0.72 to 0.57 as the number of raters was reduced from 2 to 1.

3.3. 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. Communication skills within station checklists

Table 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, 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 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.

4. Discussion and conclusion

4.1. 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-

Table 2
Summary of estimated variance component (G-Study), G-Kappa coefficient, and D Study (optimisation) analysis.

Source	*df*	Mean squares	Component	
Checklists (nested within Disciplines)	824	0.02186	0.00034	0.90%
Disciplines	3	1.4546	0.00011	0.30%
Raters	2	0.5489	0.00002	0.00%
MAAS-Global	16	12.26489	0.00362	8.60%
Rater × checklists (nested within Disciplines)	1648	0.0043	0.00025	0.60%
MAAS-Global × checklists (nested within Disciplines)	13184	0.06113	0.01553	39.00%
Disciplines × raters	6	0.3227	0.00009	0.20%
Disciplines × MAAS-Global	48	2.50588	0.00362	9.10%
Raters × MAAS-Global	32	0.97495	0.00092	2.30%
Raters × MAAS-Global × Disciplines	96	0.21141	0.00095	2.40%
Raters × MAAS-Global × checklists (nested within Disciplines)	26368	0.01455	0.01455	36.60%
G-Kappa = 0.8 (measurement design CD/RM) *				
D study (optimisation) analysis:				
Number of raters	G-Kappa			
2	0.72			
1	0.57			

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 characterised as valid items for assessing relevant CS, whereas this may not be the case for items which could not be mapped to the standard.

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)

Table 3
Summary of MAAS-Global sections and items of stations' checklists in each discipline.

Diciplines	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%)			

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 categorised 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. Generalised 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 defendable and that our assessment forms are sensitive enough to discriminate between 'good' and 'bad' performance and to measure change over time [19].

5. Limitations 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.

6. Conclusion

In the present study, station checklist items were calibrated and categorised 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 standardised against the MAAS-global, standardised

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.

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 generalisability 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.

Declaration of interest

The authors declare there is no conflict of interests.

Notes on contributors

Winny Setyonugroho, MT. is Lecturer in Health Informatics, Dept Health Informatics, Faculty of Medicine & Health Sciences, Universitas Muhammadiyah Yogyakarta, Indonesia and also a PhD student, domain Medical Informatics & Medical Education, Discipline of Medicine, School of Medicine, College of Medicine, Nursing & Health Sciences National University of Ireland Galway (supervised by Dr Thomas JB Kropmans).

Kieran M Kennedy, MB BCh BAO., MMedSci MHSc, MSc MICGP. Dr Kennedy is a Lecturer in Clinical Methods and Clinical Practice, Discipline of Medicine, School of Medicine, College of Medicine, Nursing & Health Sciences National University of Ireland Galway. He is also a practicing General Practitioner (family doctor). He has a background of specialist training and practical expertise in clinical communication skills, assessment of communication skills and the quality assurance of such assessment.

Brian Stewart, MB BCh BAO is a Lecturer in Clinical Methods and Clinical Practice, Discipline of Medicine, School of Medicine, College of Medicine, Nursing & Health Sciences National University of Ireland Galway.

Dr. Jan van Dalen is a Programme Director Master of Health Professions Education, Discipline of Skills laboratory, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

Dr Thomas JB Kropmans PhD is senior lecturer Medical Informatics & Medical Education, Domain Medical Informatics & Medical Education, Discipline of Medicine, School of Medicine, College of Medicine, Nursing & Health Sciences National University of Ireland Galway.

Acknowledgement

The lead author, Winny Setyonugroho, received a PhD scholarship from the Directorate General of Higher Education, Ministry of National Education and Culture of Indonesia.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.pec.2015.08.001.

References

[1] M. Von Fragstein, J. Silverman, A. Cushing, S. Quilligan, H. Salisbury, C. Wiskin, et al., UK consensus statement on the content of communication curricula in undergraduate medical education, Med. Educ. 42 (2008) 1100–1107, doi: http://dx.doi.org/10.1111/j 1365-2923.2008.03137.x.
[2] H. Boon, M. Stewart, Patient-physician communication assessment instruments: 1986 to 1996 in review, Patient Educ. Couns. 35 (1998) 161–176, doi:http://dx.doi.org/10.1016/S0738-3991(98)63-9.
[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] D.J. Cegala, B. Lenzmeier, S. roz, Physician communication skills training: a review of theoretical backgrounds, objectives and skills, Med. Educ. 36 (2002) 1004–1016.
[5] 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) . http://dx.doi.org/10.1016/j.pec.2015.06.004.
[6] G. Guiton, C.S. Hodgson, G. Delandshere, L. Wilkerson, Communication skills in standardized-patient assessment of final-year medical students: a psychometric study, Adv. Health Sci. Educ. 9 (2004) 179–187.
[7] Kiessling, C, Essers, G, Anvik, T, Jankowska, K., 2015. 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: 26.04.15.).
[8] E.G. Carmines, R.A. Zeller, Reliability and Validity Assessment, SAGE, 1979.
[9] J.M. Schirmer, L. Mauksch, F. Lang, M.K. Marvel, K. Zoppi, R.M. Epstein, et al., Assessing communication competence: a review of current tools, Fam. Med. 37 (2005) 184–192.
[10] D.L. Roter, J.A. Hall, D. Blanch-Hartigan, S. Larson, R.M. Frankel, Slicing it thin: new methods for brief sampling analysis using RIAS-coded medical dialogue, Patient Educ. Couns. 82 (2011) 410–419, doi:http://dx.doi.org/10.1016/j.pec.2010.11.019.
[11] 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, Adv. Health Sci. Educ. (2002) .
[12] J. Silverman, The Calgary–Cambridge guides: the teenage years, Clin. Teach. 4 (2007) 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 04.25.15.).
[14] T.J. Kropmans, B.G. O'Donovan, D. Cunningham, A.W. Murphy, G. Flaherty, D. Nestel, et al., An online management information system for objective structured clinical examinations, Comput. Inf. Sci. 5 (2011) 38, doi:http://dx.doi.org/10.5539/cis.v5n1p38.
[15] 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:http://dx.doi.org/10.3109/0142159X.2013.818635.
[16] J. Van Thiel, P. Ram, J. van Dalen, MAAS-global Manual, Maastricht Maastricht Univ, 2000 http://www.hag.unimaas.nl/Maas-global_2000/GB/MAAS-Global-2000-EN.pdf (accessed: 8.11.2012) 4–5.
[17] J. Brown, How clinical communication has become a core part of medical education in the UK, Med. Educ. 42 (2008) 271–278, doi:http://dx.doi.org/10.1111/j 1365-2923.2007.02955. x.
[18] J. Cardinet, S. Johnson, G. Pini, Applying Generalizability Theory Using EduG, Taylor & Francis, 2012.
[19] T. Kropmans, P. Dijkstra, B. Stegenga, R. Stewart, B. De, L. ont, Smallest detectable difference of maximal mouth opening in patients with painfully restricted temporomandibular joint function, Eur. J. Oral Sci. 108 (2000) 9–13.
[20] J.A. Boyko, J.N. Lavis, M. Dobbins, N.M. Souza, Reliability of a tool for measuring theory of planned behaviour constructs for use in evaluating research use in policymaking, Health Res. Policy Syst. 9 (2011) 29, doi:http://dx.doi.org/10.1186/1478-4505-9-29.
[21] K.D. Lakes, W.T. Hoyt, Applications of generalizability theory to clinical child and adolescent psychology research, J. Clin. Child Adolesc. Psychol. 38 (2009) 144–165, doi:http://dx.doi.org/10.1080/15374410802575461.
[22] A.M. Briesch, H. Swaminathan, M. Welsh, S.M. Chafouleas, Generalizability theory: a practical guide to study design, implementation, and interpretation, J. School Psychol. 52 (2014) 13–35, doi:http://dx.doi.org/10.1016/j.jsp.2013.11.008.
[23] 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:http://dx.doi.org/10.3109/0142159X.2012.703791.

[24] P.B. Andreatta, D.A. Marzano, D.S. Curran, Validity what does it mean for competency-based assessment in obstetrics and gynecology? Am. J. Obstet. Gynecol. 204 (384) (2011) e1–384, doi:http://dx.doi.org/10.1016/j.ajog.2011.01.061 e6.

[25] 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:http://dx.doi.org/10.3109/0142159X.2012.709652.

[26] M. Von Fragstein, J. Silverman, A. Cushing, S. Quilligan, H. Salisbury, C. Wiskin, et al., UK consensus statement on the content of communication curricula in undergraduate medical education, Med. Educ. 42 (2008) 1100–1107, doi:http://dx.doi.org/10.1111/j.1365-2923.2008.03137.x.