This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>

Contents lists available at SciVerse ScienceDirect

Patient Education and Counseling

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

Medical education

Optimizing the utility of communication OSCEs: Omit station-specific checklists and provide students with narrative feedback

Marc Van Nuland^{a,*}, Wim Van den Noortgate^b, Cees van der Vleuten^c, Jo Goedhuys^a^a Department of General Practice, Catholic University of Leuven, Leuven, Belgium^b Centre for Methodology of Educational Research, Catholic University of Leuven, Belgium^c Maastricht University, The Netherlands

ARTICLE INFO

Article history:

Received 12 July 2011

Received in revised form 9 December 2011

Accepted 28 December 2011

Keywords:

OSCE

Communication

Assessment

Generalizability

Utility

Feedback

Costs

ABSTRACT

Objective: To evaluate how the utility (reliability, validity, acceptability, feasibility, cost and educational impact) of a communication-OSCE was influenced by whether or not station-specific (StSp) checklists were used together with a generic instrument and whether or not narrative feedback was provided to students.

Methods: At ten stations, faculty members rated standardized patient–student interactions using the Common Ground (CG) instrument (at all stations) and StSp-checklists. Both raters and patients provided written feedback. The impact of changing the design on the various utility parameters was assessed: reliability by means of a generalizability study, cost using the Reznick model and the other utility parameters by means of a survey.

Results: Use of the generic instrument (CG) proved more reliable (G coefficient = 0.67) than using the StSp-checklists (G = 0.47) or both (G = 0.65) while there was a high correlation between both scale scores (Pearsons' r = 0.86). The cost was 6.5% higher when StSp-checklists were used and 5% higher when narrative feedback was provided.

Conclusion: The utility of a communication OSCE can be enhanced by omitting StSp-checklists and by providing narrative feedback to students.

Practice implications: The same generic assessment scale can be used in all stations of a communication OSCE. Providing feedback to students is promising but it increases the costs.

© 2012 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Mastering communication skills is acknowledged as essential in order to become a competent physician [1,2]. Many medical schools have been consistently using communication skills training in their undergraduate curricula. Training alone does not, however, guarantee better learning. One way to further enhance students' learning is to organize summative assessments, since these are known to 'drive learning' [3]. In order to assess students' communication skills we need reliable and valid assessment procedures that are suited to the stage of training. The medical education literature informs us that such assessments should definitely take case-specificity into account and that evaluations should consequently be based on multiple different clinical

situations [3,4]. In this context the Objective Structured Clinical Examination (OSCE) has become widely popular for the standardized assessment of students' skills [5,6]. A few studies within the OSCE literature have focused specifically on communication skills. These used between 6 and 18 cases with an observation time of between 60 and 360 min, were based on the use of standardized patients as raters and all used different rating scales [4,7–10] (see Table 1). They all used case-invariant (generic) rating scales and two out of the six studies also mentioned the use of case-specific rating scales [4,7] but only Cohen et al. published results about their case-specific scale [7]. This study concluded that the psychometric properties of both case-specific checklist ratings and the case-invariant rating scale were comparable but the study did not assess whether supplementing the generic scale with the case-specific scale yielded any added value. Reliability was one of the key issues in the analysis of all the studies mentioned and the published G coefficients suggested that between 7 and 18 cases (equivalent to 127 and 360 min of observation time respectively) are needed to achieve 'acceptable' reliability ($G \geq 0.80$) [4,7–10].

Van der Vleuten has called for consideration of the 'utility' of any assessment method. This means that the focus should not be

* Corresponding author at: Department of General Practice, Catholic University of Leuven, Kapucijnenvoer 33 j bus 7001, B-3000 Leuven, Belgium.
Tel.: +32 16 337 468; fax: +32 16 337 480.

E-mail addresses: marc.vannuland@med.kuleuven.be, mnuland@scarlet.be (M. Van Nuland).

Table 1
Comparison of OSCE studies with a focus on communication skills.

Study	Students	Raters	# Cases	Observation time	G-coefficient	Rating scale
Cohen et al. [7]	4th year medical students	Standardized patients	7	140 min	0.76	SIU rating scale (5 items, 5 point rating scale)
Guiton et al. [4]	Final year medical students	Standardized patients	7	127 min (6 cases of 15 min and one of 37 min)	0.65 0.82	17-item case-specific checklist 7 communication items scored on a 6-point Likert scale
Van Zanten et al. [8]	Candidates for ECFMG ^a certificate	Standardized patients	10–11	150 min	0.78	Holistic rating across 4 dimensions (4-point Likert scale)
Vu et al. [10]	Post-clerkship	Standardized patients	17–18	340–360 min	0.77–0.83	Case-invariant checklist with 2 items on communication (5-point rating scale)
Yudkowsky et al. [9]	Internal medicine and family medicine 2nd or 3rd yr residents	Standardized patients	6	60 min	0.66 0.72	CIS-scale (17 items – 5-point Likert scale) 1 global item – (5-point Likert scale)
Baig et al. [24]	International medical graduates	Standardized patients + physician examiners	14	?	0.62	Case-invariant 13 item checklist (5-point rating scale)

^a ECFMG, Educational Commission for Foreign Medical Graduates.

merely on ‘reliability’ and ‘validity’ issues, but ‘acceptability’, ‘feasibility’, ‘cost’ and ‘educational impact’ should also be considered [3,11]. For this reason the present study will address all the utility components of a newly designed communication OSCE by looking at how these are influenced by the type of assessment instrument used in the OSCE stations.

Since OSCEs are labour-intensive, time consuming and consequently costly, there is considerable interest in any way of simplifying their organization as long as their overall ‘utility’ is not jeopardized. One initial way to reduce workload and costs (thus improving feasibility) is by reducing observation time and/or the number of stations, but as we know there is a limit to this kind of modification [4,7–10]. Another way is by simplifying scoring by raters, for example by using a single validated (generic) instrument in all the stations. This offers the advantage that there is no need to evaluate the validity of the respective station-specific checklist items for each newly developed station. Furthermore, rater training is simplified if a single scale is used in all stations. This seems to be a way of enhancing feasibility and reducing costs. It remains unclear, however, whether conclusions based on the use of a generic instrument alone would be different from those based on the use of station-specific checklists because these checklists can be expected to offer added value by revealing students’ case-specific competence more effectively. In the present study we will address this issue by using the Common Ground instrument as such a generic instrument. We gained experience in the use of the validated Common Ground instrument in OSCEs through previous research [12] and found that it clearly assesses the essential tasks of competent doctor–patient communication because its development was linked to the core, generic communication skills emphasized in the Toronto and Kalamazoo consensus statements [13,14]. The Common Ground consists of seven items: six distinctive items to be scored globally (“Rapport building”/“Agenda setting”/“Information management”/“Active listening for full understanding of ideas, concerns and expectations”/“Addressing feelings”/“Reaching Common Ground”) and one last seventh item assessing the “Global total interview performance” [15].

Yet another way to enhance the ‘utility’ of an assessment is by looking for ways to enhance its educational impact for students. Our prior qualitative research had informed us that students need more feedback after OSCEs [16]. Giving such feedback can thus be expected to enhance the educational impact of the OSCE but will only enhance its ‘utility’ as long as its feasibility and cost are not jeopardized.

In summary, our primary research question concerns whether or not it is necessary to supplement a generic communication assessment instrument (like the Common Ground) with station-specific checklists in a (summative) communication OSCE. The impact of omitting checklist items on the different components of OSCE ‘utility’ will therefore be examined. We will also evaluate the impact of giving narrative feedback to students on the various components of utility.

2. Methods

2.1. Context of the research

This research was carried out at the medical faculty of the Katholieke Universiteit Leuven (University of Leuven) where there was a need in the final undergraduate year to assess medical students at the year end in a summative way: failure leads to their exclusion from General Practice Specialty Training.

2.2. Design of the communication OSCE and the assessment procedure

Station development was based on a blueprint [17] constructed from different sources: the curriculum of our own faculty, the literature on GP communication, CANmeds [1] and EURACT [2]. Based on the results of former research [4,7–10] we estimated that an OSCE with ten different stations would be sufficient for our summative purposes. The ten stations were checked against the blueprint to guarantee representative sampling across the different clinical problems and the various patients seen in GP and across the different GP communication tasks.

We constructed an assessment tool based on the Common Ground (CG) [15] supplemented by station-specific (StSp) checklist items. The Common Ground was slightly modified by using a 7-point Likert scale instead of a 5-point scale in order to gain more variance in item scores since our prior experience with the CG with the 5-point scale showed low variance [12] (We got permission from the copyright-owner F. Lang regarding this modification). At each station students were rated on all seven CG items (case-invariant) and on six StSp items (different for each station). StSp items represented key communication tasks typical for the cases described in the different scenarios. The scenarios were written in a way to guide standardized patients to play their role in a specific way and to present some verbal and non-verbal cues in a standardized way (examples of the scenarios can be obtained from the contacting author).

Blueprinting, scenarios and StSp items were critically reviewed by faculty members ($n = 4$) in order to secure content validity in advance.

In the OSCE, students were given 2 min to read the station assignment. During this time interval raters and standardized patients wrote down narrative messages intended for the students they had just seen in the same station. These messages were typed out and provided to the students together with their OSCE-scores about six weeks after the OSCE. Students had to complete each station within 7 min. There were no stations with (procedural) clinical examinations. In order to test all the students ($n = 66$) within one day for security reasons [18], we organized four sessions with two simultaneous circuits comprising the same 10 stations. In each session a group of students went through all ten stations of one of the circuits. In each station one standardized patient portrayed one of the ten roles and the student–patient interaction was observed by one rater.

Standardized patients ($n = 25$) were recruited from our group of well-trained standardized patients from final year GP vocational training OSCEs. They received their scenarios one month in advance and as they were used to portray patients they trained their part on their own.

The OSCE raters ($n = 34$) were mostly faculty members, most of whom were accustomed to rating OSCEs. Some of the others were tutors from teaching practices or GPs. They were all given an introduction to the use of the assessment tool: two weeks prior to the OSCE they received a PowerPoint tutorial by mail and later they were given a 1-h group session to discuss the tutorial and to emphasize key issues in rating the students.

2.3. How do different OSCE designs influence utility?

2.3.1. Reliability

Reliability was examined by using a two-facets (persons $\times$ stations) Generalizability Theory design [19,20]. This looked at the influence on reliability of whether station-specific checklists were used or not. Since we were also interested in the reliability of the pass/fail decision based on our standard-setting procedure, we carried out a generalizability study of the pass/fail decisions and checked for the effect of supplementing the CG scale with the StSp scale on the respective generalizability coefficients. These coefficients (being a fraction between 0 and 1) integrate the discriminating ability of the test and the reproducibility of the results [20]. We used the borderline regression method to calculate

the pass/fail threshold (for technical details about this method see the study of Kramer et al. [21]).

2.3.2. Validity, acceptability and educational impact

We checked the concurrent validity of both scales by computing Pearson's correlation coefficients. Content and face validity were further assessed by means of a survey among the participating students (about four months after the OSCE). The survey was constructed in cooperation with a university team specializing in such surveys and some of the questions were inspired by the results of our prior focus group research [16]. The survey also addressed acceptability and educational impact. The utility components addressed in the survey are not expected to be influenced by whether or not station-specific rating scales are used, but they are informative in relation to giving feedback.

2.3.3. Feasibility and cost

Feasibility was assessed by interviewing both standardized patients and raters in group sessions after the OSCE. To calculate and discuss the costs we relied on Reznick's model [22], which identifies four different phases: the development, production, administration and post-examination analysis-reporting phase. We only reported on 'high-end costs' (see Reznick) because these are indicative of the real cost whereas 'low-end costs' assume that costs in relation to faculty members are already accounted for in the faculty budget. We checked for differences in terms of costs associated with whether or not station-specific ratings were used at the different stages. We also estimated the cost reduction if the OSCE (modified to safeguard "security" [18]) was repeated in subsequent years. We also investigated the extent to which the cost would be affected if the number of students exceeded the capacity of the current OSCE. We further evaluated the impact on costs of giving feedback messages to students.

2.4. Data and analysis

To calculate station scores per student for both the Common Ground and Station Specific rating scales, the mean item score was calculated. A combined score (CG + StSp) was calculated by averaging the two scores, resulting in a score between 1 and 5. The standard error of measurement from the generalizability study was used to compute 95% confidence intervals for the calculated borderline score and its lower margin was set as a more lenient fail threshold.

Table 2

Descriptive statistics of students' total station scores ($n = 63$ students).

Stations	Common Ground total station scores (scale 1–7)			StSp total station scores (scale 1–4)			Pearson's correlation between students' CG and StSp*	Combined station score (scale 1–5)		
	Mean	SD	Range	Mean	SD	Range		Mean	SD	Range
1. First prescription contraceptive pill	4.3	1.34	1.86–7.00	2.66	0.60	1.17–4.00	0.72	3.48	0.91	1.58–5.5
2. Motivating a diabetic patient	4.52	0.98	1.57–6.71	3.09	0.5	1.83–4.00	0.74	3.81	0.7	1.87–5.36
3. Man with unstable angina	3.63	1.04	1.29–5.71	2.75	0.61	1.00–4.00	0.62	3.19	0.74	1.64–4.86
4. Unplanned pregnancy	4.56	1.15	1.57–6.86	2.62	0.63	1.17–4.00	0.77	3.59	0.84	1.52–5.43
5. Low back pain	4.89	0.79	3.29–6.57	2.86	0.44	1.50–3.67	0.66	3.88	0.57	2.64–5.09
6. Positive mammography	4.27	1.15	1.71–6.43	2.63	0.52	1.17–3.50	0.76	3.45	0.79	1.51–4.88
7. Depressive symptoms	4.76	0.92	1.86–6.29	2.32	0.53	1.00–3.67	0.43	3.54	0.64	1.68–4.56
8. Multiple reasons for encounter	4.54	1.03	2.57–7.00	2.72	0.56	1.33–4.00	0.82	3.63	0.77	1.95–5.33
9. Coping with a partner suffering from dementia	4.82	1.13	2.00–6.75	2.85	0.79	1.00–4.00	0.82	3.83	0.92	1.64–5.21
10. Student with irritable bowel symptoms	4.46	1.26	1.57–6.86	2.91	0.61	1.67–4.00	0.70	3.69	0.87	1.70–5.26
Total OSCE score	4.46	0.54	2.99–5.59	2.74	0.24	2.14–3.22	0.86	3.61	0.38	2.56–4.35

* All correlation coefficients have a p -value < 0.0001 .

2.5. Statistical software

We used SAS software (version 9.2) for all analyses except the generalizability analyses, for which GENOVA software was used.

3. Results

3.1. Descriptive statistics on performance of students

The results are based on the scores of 63 participating students. Three students were excluded from the analysis due to missing scores. The total student station scores (Table 2) appeared to vary between students and between stations with mean CG scores between 3.63 and 4.89 and variances (SD^2) fluctuating between 0.96 and 1.80, with mean StSp scores between 2.32 and 3.09 and variances between 0.19 and 0.62. Combined scores were between 3.19 and 3.88 and variances were between 0.32 and 0.85. Table 2 also illustrates the concurrent validity of the Common Ground and the Station Specific scale by providing the respective Pearson's correlation coefficients. Correlations range from moderate to strong for most stations, except the weaker correlation in station 7.

3.2. The impact of changing the OSCE design on reliability

Table 3 presents how estimates for the different variance components changed depending on which rating scale was used. It appears that the variance component for students (P) is higher for the Common Ground than for the station-specific checklist, which indicates better discriminating power when the Common Ground is used. Differences in station difficulty are reflected by the station

Table 3

Variance components from generalizability study.

	P	S	PS
CG	15.33%	8.81%	75.86%
StSp	7.44%	9.93%	82.63%
Combined	14.86%	5.44%	79.69%

variance components, which accounted for between 5.44% and 9.93% of the total variance depending on which scale was used.

The D-study showed a (norm-referenced) Generalizability coefficient for the combined scale of 0.65 when 10 stations are used, reaching 0.80 if 20 stations were used. For the CG and StSp scales the G coefficients for 10 stations were 0.67 and 0.47 respectively (see Table 4). In conclusion, the combined scale reveals a G-coefficient which is comparable to but lower than for CG alone.

In a separate G and D study we investigated the generalizability of the decision based on our borderline-based standard setting procedure (Table 4). It appeared that for both the combined and CG scales this generalizability coefficient was 0.92 with 10 stations. In other words, with regard to the generalizability of pass/fail decisions there seems to be no need to supplement the CG scale by using StSp items.

3.3. Impact of changing OSCE designs on students' views on validity, acceptability and educational impact

A total of 55 students participated in the survey. Students seem to appreciate being given narrative feedback (mean = 4.82 (SD:

Table 4
Generalizability D studies.

	Number of stations	6	8	10	12	14	16	18	20
Common Ground	G-coefficient	0.55	0.62	0.67	0.71	0.74	0.76	0.78	0.80
	SEM	0.41	0.35	0.31	0.29	0.27	0.25	0.23	0.22
Station specific	G-coefficient	0.35	0.42	0.47	0.52	0.56	0.59	0.62	0.64
	SEM	0.23	0.20	0.18	0.16	0.15	0.14	0.13	0.13
Combined scale	G-coefficient	0.53	0.60	0.65	0.69	0.72	0.75	0.77	0.79
	SEM	0.29	0.25	0.23	0.21	0.19	0.18	0.17	0.16
Decision with CG scale	G-coefficient	0.87	0.90	0.92	0.93	0.94	0.95	0.95	0.96
Decision with StSp	G-coefficient	0.84	0.88	0.90	0.92	0.93	0.94	0.94	0.95
Decision with combined scale	G-coefficient	0.88	0.91	0.92	0.94	0.94	0.95	0.96	0.96

The D studies are based on our experiment with ten stations (highlighted column).

Table 5
Cost of communication OSCE in first year and subsequent years.

Reznick phases	Total cost in first year			Total cost in subsequent years (capacity 80 students)			Total cost in subsequent years (capacity 120 students)		
	Combined scale + feedback	CG only + feedback	CG only + no feedback	Combined scale + feedback	CG only + feedback	CG only + no feedback	Combined scale + feedback	CG only + feedback	CG only + no feedback
1. Development	€7200 31.68%	€6200 29.07%	€6200 30.50%	€2800 15.97%	€ 2500 14.68%	€ 2500 15.50%	€ 2800 11.62%	€ 2500 10.60%	€ 2500 11.02%
2. Production	€2110 9.28%	€2110 9.89%	€2110 10.38%	€2110 12.04%	€ 2110 12.39%	€ 2110 13.08%	€ 2762 11.47%	€ 2762 11.71%	€ 2762 12.17%
3. Administration	€10,619 46.72%	€10,619 49.79%	€10,619 52.24%	€10,619 60.58%	€ 10,619 62.36%	€ 10,619 65.84%	€ 15,929 66.12%	€ 15,929 67.52%	€ 15,929 70.20%
4. Post-examination	€2800 12.32%	€2400 11.25%	€1400 6.89%	€2000 11.41%	€ 1800 10.57%	€ 900 5.58%	€ 2600 10.79%	€ 2400 10.17%	€ 1500 6.61%
Total costs	€22,729	€21,329	€20,329	€17,529	€ 17,029	€ 16,129	€ 24,091	€ 23,591	€ 22,690
Costs/student if 66 students	€344	€323	€308	€266	€ 258	€ 244			
Costs/student if 80 students	€284	€267	€254	€219	€ 213	€ 202			
Costs/student if 100 students							€ 241	€ 236	€ 227
Costs/student if 120 students							€ 201	€ 197	€ 189

1.11) on a Likert scale from 1 to 6) but more importantly students seem to learn more from the narrative messages (mean 3.60 (SD: 1.41)) than from their numerical scores (mean 3.20 (SD: 1.34)) on the rating scale. Students also agreed that the OSCE is an appropriate way to assess their communication skills (mean = 4.15 (SD: 1.03)), agreed on its content validity (mean = 4.93 (SD: 0.74)) and on the representativeness of the 10 stations for GP consultations (mean = 4.56 (SD: 0.23)).

3.4. Impact of changing OSCE design on feasibility and costs

When evaluating feasibility and costs there is a difference between the initial implementation of the OSCE and use of the OSCE in subsequent years, taking into account the need to make certain modifications in order to safeguard security [17,22] (see Table 5). In our experience the investment of time – and consequently the cost – of the development phase was more than twice as high in the first year. On the other hand, production costs (secretarial support and materials) and administration costs remain the same during subsequent years. In the post-examination phase, the cost of analysis is reduced (by 50%) in subsequent years whereas the cost of secretarial support remains the same. Supplementing the Common Ground scale with StSp checklists increased costs by up to 6.5%, mainly by increasing the development and post-examination workload. Providing students with feedback resulted in a cost increase of about 5%, mainly related to extra secretarial workload (transcription of feedback). Using both StSp checklists and narrative feedback may therefore increase the cost by up to 12%.

Table 5 further illustrates that the number of participating students in relation to the capacity of the OSCE has a substantial impact on the price per student. The maximum capacity of the present OSCE is 80 students. A larger number of students would demand more time of simulated patients and observers (necessitating an additional OSCE session) and would increase the secretarial workload.

4. Discussion and conclusion

4.1. Discussion

The main finding of our study is that the assessment of communication skills using a 10-station communication OSCE can rely solely on using a generic scale such as the Common Ground, rendering the construction and use of station-specific checklists redundant. The fact that there is no need to supplement the CG with StSp items is illustrated by the results of the generalizability decision study and by the high correlations between CG and StSp scales, although these high correlations may, of course, be attributable to a halo effect. The *G* coefficients that we obtained are comparable to those mentioned in other studies (see Table 1), if observation time is also taken into account. By omitting the station-specific checklists it would be possible to reduce the time spent creating and validating the checklists, time spent training/instructing observers and the time and intellectual exertion devoted to observation, meaning that it would improve acceptability and feasibility, resulting according to our calculations in a

cost reduction of about 6%. McKinley et al. came to similar conclusions in constructing a generic (holistic) rating scale for assessing procedural skills, potentially rendering context-specific scales redundant, except that they did not include cost calculations [23]. On the other hand the study by Baig et al. [24] regarding communication skills assessments may seem to contradict our conclusion when it states that “a generic instrument may not be useful for all cases (stations)”, a conclusion based on their observation that the 14 items in a case-invariant checklist were rated inconsistently across stations for the same student. This different conclusion, however, illustrates that whether or not a generic instrument is ‘useful’ depends on the purpose of the assessment. If one wishes to evaluate overall competence in communication, as we did, a generic instrument appears to be appropriate. On the other hand, if the aim is to evaluate specific context-related issues, a case-specific checklist will be more appropriate. It is interesting to note from the variance components that the global judgment represented by the CG instrument leads to more person variance than the station-specific items. Apparently global judgments tend to pick up aspects that are more generalizable across different contexts. The same has also been found through prior OSCE research [25]. Global judgments tap into something that is less content-specific and more transferable, while station-specific items are more anchored to the specific situation and therefore induce more contextual variability.

Regarding the cost of OSCEs, comparisons between different studies must be interpreted with a similar scrutiny. An older review of OSCE costs referred to high-end costs per student from USD 200 (€137) to USD 1020 (€700) [26]. In a more recent review this range was between USD 6.90 (€4.8) and USD 1020 (€700) but this also included studies only reporting low-end costs [27]. The cost analysis presented by us reveals substantial differences in the cost per student between the first and subsequent uses of the OSCE, whether or not the OSCE is functioning at maximum capacity, whether or not supplementary case-specific checklists are used and whether or not narrative feedback is given to students. Our costs per student (€189–344 for 10 stations – 70 min of observation time) are approximately comparable with the costs mentioned in three recent OSCE studies [28–30] taking into account the number of stations and whether or not all costs were included (see Table 6).

Giving narrative feedback to students seems to be highly appreciated by students and may have a formative effect, an opportunity which is often missed when only numerical scores are provided. Such quality improvements come at a price, however (a 5% increase in the cost). Based on our results, we would rather support reducing costs by using the generic scale without supplementing it with StSp checklists, but to give feedback to students in order to improve the educational impact in spite of the extra costs. This is illustrative of how the ‘utility’ concept can serve as a guide to ways of ‘improving’ assessments taking into account the priorities defined by faculty policy.

4.1.1. Strengths and weaknesses of our study

To our knowledge this is the first study to investigate whether the assessment of communication in an OSCE format can rely solely on a generic assessment instrument by addressing the

Table 6
Published OSCE costs per student in comparable studies between 2004 and 2011.

Study	Number of stations	Observation time (min)	Number of students	Cost per student	Remarks
Kelly and Murphy [28]	5	25	71	€102	Unclear if development and post-examination costs were included
Varkey et al. [29]	8	135	9	€176 (\$255)	Did not include 45 h faculty time for development of the OSCE
Palese et al. [30]	6	60	80	€145	High-end costs

different components of 'utility' as defined by van der Vleuten [3,11]. By selecting the Common Ground we were using a communication assessment tool that had already been validated and was conceived on the basis of a consensus on the essential elements of doctor–patient communication [13,14]. It seems improbable that by changing the original 5-point scale to a 7-point scale the performance of the Common Ground was affected. The results of our Generalizability study are providing arguments for the reliability of the CG using the 7-point scale. Furthermore we have been computing Cronbach's alpha's for each station and found that the CG items scored communication in each station in a consistent way (Alpha's ranged between 0.85 and 0.93 with a mean of 0.92 meaning high internal consistency).

A weakness of our study is that we investigated the educational impact of giving feedback solely by means of two survey questions to assess the subjective appraisal of students. We did not investigate in depth whether students really 'learned' from the feedback messages. Another weakness may be that we did not systematically ask raters about the perceived extra burden when they were required to rate both the Common Ground and the station-specific checklist items. Despite these weaknesses we believe that our results give an indication of the investments in terms of time and cost that are required to observe medical trainees and obtain a clear and fair view of their competence in communication.

4.1.2. Further research

It is clear that our research into a communication OSCE would benefit from being replicated in other localities or different contexts and with different designs (e.g. with a larger number of raters to tease out rater-dependent and case-dependent sources of variance in relation to specific and generic rating scales). It would also be interesting to investigate the predictive validity of our OSCE by checking which scales provide a better predictor of actual performance in day-to-day communication in subsequent years. Another point of interest may be to evaluate the use of the Common Ground instrument in integrated clinical skills OSCEs to assess the aspect of communication in these stations since it cannot be taken for granted that our results are transferable to such OSCEs. Finally, new research could also further explore the educational impact of communication OSCEs, e.g. by exploring whether they generate stimuli to address the specific learning needs of individual students.

4.2. Conclusion

In communication OSCEs it is possible to rely on the use of a single generic assessment scale (such as the Common Ground) in all stations without the need to supplement them with station-specific checklists. This has a positive impact on the overall 'utility' of the assessment because it enhances feasibility and cost without jeopardizing the other components of the 'utility' concept: reliability, validity, acceptability and educational impact. As regards the educational impact, a positive effect can be expected from giving OSCE station related narrative feedback to students. As well as improving the educational impact this also increases acceptability, but conversely it does result in some additional costs. Faculty policy must set priorities to guide such adjustments.

4.3. Practice implications

The utility of a communication OSCE is enhanced by omitting STSp checklists and by using the same generic assessment scale in all stations. Providing narrative feedback to students is promising in improving the educational impact of OSCEs but it increases the costs.

Acknowledgments

Our special thanks go to Ron Hoogenboom for conducting the Genova analysis, to Steven Huyghe and Hilde Creten from DUO (Dienst Universitair Onderwijs from KULeuven) for their help with the survey and to Marina Devis for her secretarial support.

References

- [1] Frank J, editor. The CanMeds 2005 physician competency framework. Better standards. Better physicians. Better care. 2008. p. 1–40, http://rcpsc.medical.org/canmeds/CanMEDS2005/CanMEDS2005_e.pdf.
- [2] Heyrman J, editor. EURACT educational agenda. 2005. p. 1–48, <http://www.euract.org/pdf/agenda.pdf>.
- [3] van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1996;1:41–67.
- [4] 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 Theory Pract* 2004;9:179–87.
- [5] Newble D. Techniques for measuring clinical competence: objective structured clinical examinations. *Med Educ* 2004;38:199–203.
- [6] Patricio M, Juliao M, Fareira F, Young M, Norman G, Vaz CA. A comprehensive checklist for reporting the use of OSCEs. *Med Teach* 2009;31:112–24.
- [7] Cohen DS, Collier JA, Marcy MS, Fried ED, Swartz MH. Psychometric properties of a standardized-patient checklist and rating-scale form used to assess interpersonal and communication skills. *Acad Med* 1996;71:S87–9.
- [8] van Zanten M, Boulet JR, McKinley D. Using standardized patients to assess the interpersonal skills of physicians: six years' experience with a high-stakes certification examination. *Health Commun* 2007;22:195–205.
- [9] Yudkowsky R, Downing SM, Sandlow LJ. Developing an institution-based assessment of resident communication and interpersonal skills. *Acad Med* 2006;81:1115–22.
- [10] Vu NV, Marcy ML, Verhulst SJ, Barrows HS. Generalizability of standardized patients' satisfaction ratings of their clinical encounter with fourth-year medical students. *Acad Med* 1990;65:S29–30.
- [11] van der Vleuten C, Schuwirth LW. Assessing professional competence: from methods to programmes. *Med Educ* 2005;39:309–17.
- [12] Van Nuland M, Van Den Noortgate W, Degryse J, Goedhuys J. Comparison of two instruments for assessing communication skills in a general practice objective structured clinical examination. *Med Educ* 2007;41:676–83.
- [13] Simpson M, Buckman R, Stewart M, Maguire P, Lipkin M, Novack D, et al. Doctor–patient communication: the Toronto consensus statement. *Br Med J* 1991;303:1385–7.
- [14] Makoul G. Essential elements of communication in medical encounters: the Kalamazoo consensus statement. *Acad Med* 2001;76:390–3.
- [15] Lang F, McCord R, Harvill L, Anderson DS. Communication assessment using the Common Ground instrument: psychometric properties. *Fam Med* 2004;36:189–98.
- [16] Van Nuland M, Thijs G, Van Royen P, Van den Noortgate W, Goedhuys J. Vocational trainees' views and experiences regarding the learning and teaching of communication skills in General Practice. *Patient Educ Couns* 2010;78:65–71.
- [17] Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ* 2002;36:800–4.
- [18] Swanson D, Clauser B, Case S. Clinical skills assessment with standardized patients in high-stakes tests: a framework for thinking about score precision, equating, and security. *Adv Health Sci Educ* 1999;4:67–106.
- [19] Crossley J, Russell J, Jolly B, Ricketts C, Roberts C, Schuwirth L, et al. 'I'm pickin' up good regressions': the governance of generalisability analyses. *Med Educ* 2007;41:926–34.
- [20] Crossley J, Davies H, Humphris G, Jolly B. Generalisability: a key to unlock professional assessment. *Med Educ* 2002;36:972–8.
- [21] Kramer A, Muijtjens A, Jansen K, Dusman H, Tan L, van der Vleuten C. Comparison of a rational and an empirical standard setting procedure for an OSCE. Objective structured clinical examinations. *Med Educ* 2003;37:132–9.
- [22] Reznick RK, Smee S, Baumber JS, Cohen R, Rothman A, Blackmore D, et al. Guidelines for estimating the real cost of an objective structured clinical examination. *Acad Med* 1993;68:513–7.
- [23] McKinley RK, Strand J, Gray T, Schuwirth L, Alun-Jones T, Miller H. Development of a tool to support holistic generic assessment of clinical procedure skills. *Med Educ* 2008;42:619–27.
- [24] Baig L, Violato C, Crutcher R. Assessing clinical communication skills in physicians: are the skills context specific or generalizable? *BMC Med Educ* 2009;9:22.
- [25] Govaerts MJ, Van der Vleuten CP, Schuwirth LW. Optimising the reproducibility of a performance-based assessment test in midwifery education. *Adv Health Sci Educ Theory Pract* 2002;7:133–45.
- [26] Carpenter JL. Cost analysis of objective structured clinical examinations. *Acad Med* 1995;70:828–33.

- [27] Walsh M, Bailey PH, Koren I. Objective structured clinical evaluation of clinical competence: an integrative review. *J Adv Nurs* 2009;65:1584–95.
- [28] Kelly M, Murphy A. An evaluation of the cost of designing, delivering and assessing an undergraduate communication skills module. *Med Teach* 2004;26:610–4.
- [29] Varkey P, Natt N, Lesnick T, Downing S, Yudkowsky R. Validity evidence for an OSCE to assess competency in systems-based practice and practice-based learning and improvement: a preliminary investigation. *Acad Med* 2008;83:775–80.
- [30] Palese A, Bulfone G, Venturato E, Urli N, Bulfone T, Zanini A, et al. The cost of the objective structured clinical examination on an Italian nursing bachelor's degree course. *Nurse Educ Today* 2011, <http://www.nurseeducationtoday.com/article/abstracts?terms1=palese&terms2=&terms3=&terms4>.