Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.JournalofSurgicalResearch.com

Standardized and quality-assured video-recorded examination in undergraduate education: informed consent prior to surgery☆

Christoph Kiehl,[a] *Anne Simmenroth-Nayda,*[b] *Yvonne Goerlich,*[c] *Andrew Entwistle,*[d] *Sarah Schiekirka,*[e] *B. Michael Ghadimi,*[a] *Tobias Raupach,*[f] *and Sarah Koenig*[a,*]

[a] *Department of General, Visceral and Paediatric Surgery, University Medical Centre Goettingen, Goettingen, Germany*
[b] *Department of General Practice, University Medical Centre Goettingen, Goettingen, Germany*
[c] *Student Training Centre of Clinical Practice and Simulation, University Medical Centre Goettingen, Goettingen, Germany*
[d] *Department of Genetic Epidemiology, University Medical Centre Goettingen, Goettingen, Germany*
[e] *Office of the Dean of Studies, University Medical Centre Goettingen, Goettingen, Germany*
[f] *Department of Cardiology and Pneumonology, University Medical Centre Goettingen, Goettingen, Germany*

ARTICLE INFO

Article history:
Received 8 December 2013
Received in revised form
23 January 2014
Accepted 24 January 2014
Available online 30 January 2014

Keywords:
Medical education
Assessment
OSCE
Informed consent
Communication skills
Simulated patients
Undergraduates
Quality assurance
Reliability of checklists
Interrater agreement

ABSTRACT

Background: Communication skills combined with specialized knowledge are fundamental to the doctor–patient relationship in surgery. During a single-station video-recorded objective structured clinical examination (VOSCE), students were tasked with obtaining informed consent. Our aim was to develop a standardized and quality-assured assessment method in undergraduate education.

Methods: One hundred fifty-five students in their fifth year of medical school (78 videos) participated in a summative VOSCE within the framework of the teaching module "Operative Medicine." They prepared for three clinical scenarios and the surgical procedures involved. The examination comprised participants having to obtain informed consent from simulated patients, video recording their performance. Students were assessed by two independent raters, the background of one of whom was nonsurgical. Results were statistically tested using SPSS.

Results: Students' scores were all beyond the pass mark of 70%, averaging 91.0% (±4.0%), 88.4% (±4.4%), and 87.0% (±4.7%) for the appendectomy, cholecystectomy, and inguinal hernia repair checklist, respectively. Most items (68%–89% of the checklists) were found to have fair to excellent discrimination values. Cronbach's α values ranged between 0.565 and 0.605 for the individual checklists. Interrater agreement was strong (Pearson correlation coefficient = 0.80, $P < 0.01$; intraclass correlation coefficient 2.1 = 0.78).

Conclusions: The VOSCE is both feasible and reliable as a method of assessing student communication skills and the application of clinical knowledge while obtaining informed

* *Corresponding author*. Department of General, Visceral and Paediatric Surgery, University Medical Centre Goettingen, Robert-Koch-Strasse 40, D-37075 Goettingen, Germany. Tel.: +49 551 398977; fax: +49 551 396109.
E-mail address: skoenig1@gwdg.de (S. Koenig).

http://dx.doi.org/10.1016/j.jss.2014.01.048

consent in surgery. This method is efficient (flexible rating outside normal working hours possible with reductions in administrative load) and may be used for high-stakes evaluation of student performance.

1. Introduction

Communication skills are considered to be a core proficiency and are crucial to professionalism in medical practice, including successful outcomes in patient care [1]. Most medical schools include communication skills training in their undergraduate curricula. However, training alone does not guarantee better learning. One way of further enhancing study is to organize summative assessments because these are known to "drive learning" [2]. To assess students' skills, reliable and valid assessment procedures are needed that are suited to the stage of training. In this context, the objective structured clinical examination (OSCE) has become popular in the assessment of clinical performance in a wide range of settings [3,4].

Only a few studies within the OSCE literature have focused on how best to teach and assess communication skills with respect to surgical education in undergraduates. Published examples typically focus on delivering "bad news" to patients [5]. Indeed, a surgery-specific communication OSCE was established in the context of end-of-life communication training during surgical clerkship [6] or in the context of formative assessment of postgraduate clinical training involving six surgical scenarios for common communication tasks and interpersonal skills [4]. However, there are no satisfactory reports describing how to implement a quality-assured OSCE centered on undergraduates obtaining informed consent. For medical students in particular, this competency is often regarded as multifaceted and complex, as a properly conducted surgical informed consent process needs to provide patients with the means to authorize an invasive procedure with full comprehension of the relevant information including involved risks. Thus, obtaining informed consent comprises a multitude of educational objectives (the third level of Miller's pyramid, "shows how" [7]): cognitive and communication skills, as well as professionalism focusing on the specific needs of the patient [8]. Of note, medical students need to practice relevant clinical skills up to a routine level under supervision. In this context, the OSCE format appears the most suitable to assess the multitude of combined learning objectives associated with the task of obtaining informed consent. The OSCE provides important elements of quality assurance (metrics), as both examiners and simulated patients (SPs) can be trained and virtual clinical scenarios enable reproducibility [9,10].

The educational environment in surgery is known to be plagued by interfering clinical duties (e.g., theatre schedules, emergencies). Therefore, a video-recorded OSCE (VOSCE) with time-shifted rating may prove to be an efficient substitute for real-time live assessment. Of course, filming is not an entirely novel concept in this context. Vivekananda-Schmidt *et al.* [11] implemented a VOSCE to assess musculoskeletal examination skills in undergraduate students. Video recording of a communication session was recently reported as a means of assessing students during the preclinical phase [12,13].

In our study, we considered the filming element of the VOSCE as being indispensable to the appraisal of an entire semester cohort. Our aim was to develop and implement a single-station VOSCE during the fifth year of a German medical school centered on obtaining informed consent. Our study outlines the feasibility of the VOSCE in undergraduate education in surgery and comments on the benefits of time-shifted rating by means of video. The format of an OSCE was used for high-stakes testing, as it was essential to demonstrate quality assurance allowing fair and rigorous decision making with respect to candidates. In particular, we compared student performance in the three scenarios and analyzed the reliability and internal consistency of the checklists. For further improvements in quality, we investigated the extent of agreement between two trained raters, the background of one of whom lay outside the field of surgery.

2. Methods

2.1. *Setting and participants*

We designed a cross-sectional study with data acquisition from a summative examination. The study ran during the 5 weeks teaching module "Operative Medicine" during the summer semester of the fifth year (academic year 2010/2011) of the degree of human medicine at the University Medical Centre Goettingen, Germany (UMG). Like most German medical schools, the UMG offers a 6 years curriculum comprising two preclinical and three clinical years, followed by a practical year. The clinical curriculum is modular in structure; the sequence of modules is identical for all students. During the module Operative Medicine, knowledge and skills are recapitulated in various surgical specialties (visceral, orthopedic/trauma, and thorax/heart/lung) through emphasis on clinical decision making and patient management. In preparation (longitudinal curriculum), students are required to take a course in communication skills (with SPs) at the beginning of the third year. Furthermore, they also attend a 1 week clinical skills in surgery block during the fourth year, which includes teaching during patient encounters on the ward.

All 155 students enrolled in the teaching module participated. The average age was 25.7 ± 2.1 years. A total of 53.8% of the participating students were females and 46.2% were males. Following consultation with the University Ethics Committee, approval was not required for this type of educational study. Written consent was obtained from the students for the filming and for use of the data within the framework of the current study.

Students were requested to form pairs with a partner of their choice to prepare for and undergo the examination. Following a specific introductory lecture on the legal aspects of informed consent, as well as on the specific content and course of events, students prepared with information on all three scenarios (acute appendicitis, cholecystolithiasis, and inguinal hernia) and the surgical procedures involved (laparoscopic appendectomy, laparoscopic cholecystectomy, and open hernia repair with alloplastic mesh [Lichtenstein procedure]).

Students were provided with a manual including the consent forms, technical information on the procedures, and textbook summaries of the subjects involved. During self-study as officially allocated time (6 hours on their timetable), students reviewed this information and had to consider the structure and course of the patient interview, including communication skills and content related to the clinical cases and procedures. Students were also asked to practice obtaining informed consent from their peers prior to examination.

2.2. SPs and scenarios

The SPs were selected from a group of professional actors who regularly perform for medical training purposes. Written consent was obtained from the SPs for the filming, following which they were specifically trained using the three scenarios developed for the VOSCE. They were prepared with five relevant questions to ensure interaction with the medical students during assessment. Content validity was addressed by having the SP scenarios, roles, and checklists for rating written by an experienced surgeon trained in medical education issues and familiar with the "Goettingen Catalogue of Learning Objectives" [14].

2.3. Examination

The VOSCE was carried out in the style of an OSCE consisting of one station only. The VOSCE took place on five dates during the teaching module (every Thursday afternoon); student pairs were allocated randomly. Each student pair had to hand in one video for assessment, 78 videos in total were collected as data files and finally assessed by both raters.

On their day of examination, the student pairs were informed of the clinical scenario and then went on to obtain informed consent from a SP. The student pairs were given 30 minutes to perform two interviews (one each) with each interview lasting no longer than 10 minutes. These interviews were recorded on a tripod-mounted digital video camera equipped with an external microphone. Following initial instruction by student peers, the camera was operated by the candidates themselves. The interviews were recorded to digital media, and files were transferred to an external hard drive. The two students then moved on to another room, viewed their videos, discussed their performance on peer level, and finally selected one filmed interview for final assessment.

2.4. Checklists

Participants were allocated to the three checklists randomly. We recorded background data on informed consent talks, which students might have performed during voluntary clerkships, to ensure that there were no confounders among the three checklists.

Each checklist comprised a total of 26 items. Part A (communication) assessed verbal and nonverbal skills and comprised seven items for global rating on a 6-point Likert scale, on which 6 is excellent and 1 is unsatisfactory. Part B (content) specified the indication for surgery, choice of procedure, general and specific risks, and postoperative treatment/follow-up. Part B included two items as described for Part A and 17 items on a binary scale (2 or 1 for "done" or "not done", respectively). The scores of individual items were summed; the weighting of Part A (maximum of 42 points) to Part B (maximum of 46 points) was set as 3:7 (Part A = 30% and Part B = 70%). For scoring and visualization of data, absolute scores were converted into percentages. The total minimum percentage pass rate was set to 70%.

2.5. Raters

Examination performance was determined by two independent raters. Both raters were third-year residents. Rater 1 was a surgical resident in specialty training in general and visceral surgery. Rater 2 was a qualified dentist with three years of clinical experience. Both raters had no prior experience in scoring OSCEs but had been given instructions and training prior to the examination. All candidates had their chosen film assessed by both raters individually. Rater assessment took place out of normal working hours, for which the raters received financial compensation.

2.6. Statistical analysis

Statistical testing was performed using IBM SPSS Statistics version 19. Absolute score point values were converted to percentage scores. Means, medians, standard deviation, and confidence intervals were calculated for the scores. A sample size of 24–28 videos per checklist was considered mandatory for descriptive statistics.

Item analysis within classic test theory relies on two statistics: the *P*-value (item difficulty) and the *r*-value (item discrimination). Item difficulty was defined as proportion on a scale of 0–1 of students answering the item correctly, the value 1 indicating that all candidates were successful on the item. The item discrimination, otherwise referred to as corrected item–total correlation, is a useful measure of item quality whenever the purpose of a test is to produce a spread of scores, reflecting differences in student performance. It indicates the extent to which success on an item corresponds to success on the whole test [15,16].

Cronbach's α was used as a measure of internal consistency, that is, how closely related a set of items are as a group. In other terms, Cronbach's α is a function of the extent to which items in a test have high commonalities and thus low uniqueness [17]. A "high" value of alpha is often used as evidence that the items measure an underlying (or latent) construct [18].

Any association between the individual scoring by each rater was assessed by intraclass correlation coefficient (ICC) and Pearson contingency coefficient (PCC). ICC is used for quantitative measurements made on units that are organized into groups. It describes how strongly units in the same group resemble each other [19]. PCC is a measure of the linear correlation (dependence) between two variables *X* and *Y*, giving a value between +1 and −1 inclusive, where 1 is total positive correlation, 0 is no correlation, and −1 is total negative correlation [20].

One-way analysis of variance (ANOVA) and Scheffé's method were performed to analyze variance among the checklists. ANOVA provides a statistical test of whether the means of several groups are equal, and therefore generalizes the *t*-test to more than two groups [21]. Scheffé's method is a test for adjusting significance levels in a linear regression analysis to account for multiple comparisons [22].

3. Results

3.1. Evaluation of checklists and items

Internal consistency was assessed for each checklist and their respective parts (Table 1). In our study, the reliability (Cronbach's α) of the individual checklists ranged from 0.565–0.605 for total values. It must be stressed that these values are high for a single-station design. The highly homogeneous nature of student performance in Part A led to a low reliability in all three checklists (0.201–0.384). In contrast, student performance in Part B proved much more heterogeneous. As a direct result, Part B was determined to be more reliable than Part A, with increased reliability ranging from moderate to substantial (0.583–0.623).

As Part B of the checklists assessed the content of the informed consent interview, descriptive statistics of checklist items were determined in more detail. Item difficulty (*P*) and item discrimination (*r*) of scoring were evaluated (Table 2). Statistical analysis demonstrated that all the three checklists had good quality. A desirable item difficulty ($P = 0.4–0.8$) was determined for most items (47%, 63%, and 58% of the appendectomy, cholecystectomy, and hernia repair checklist, respectively). The item discrimination ($r > 0.2$) was fair to high in 89%, 84%, and 68%, respectively. Combining these two criteria, the items "diagnosis/indication," "choice of procedure," and "injury to neighboring organs" on the appendectomy checklist contributed to the high quality of the checklist. On the cholecystectomy checklist, the items "conversion to open surgery," "scarring," "adhesions/bowel obstruction," "incisional hernia," and "aerodermectasia" fulfilled these criteria. The hernia repair checklist even contained seven items attributing to the high quality: "diagnosis/indication," "thrombosis/embolism," "scarring," "incisional hernia," "injury/constriction of inguinal nerves," "chronic inguinal pain," and "return to normal diet/ambulation." Our evaluation of single items enabled the assessment of student performance on the level of specific learning objectives. On the cholecystectomy checklist, for example, the item "conversion to open surgery" was of good quality, with 71% of students explaining the content of this item correctly to the SPs (item difficulty 0.71) and with an item discrimination of 0.253 (classified as "fair" to distinguish between knowledgeable students and those who are not). In contrast, the item "positioning injury" was of poor quality with only 17% of student explanations proving correct as well as a low discrimination (0.038), implying that candidates performing well in the rest of the test performed poorly on this item and vice versa. A negative discrimination index indicates that the item is measuring something other than the rest of test (e.g., "extending the scope of surgery" with $r = -0.117$).

Table 1 – Cronbach's α of the three checklists.

Appendectomy	Part A	0.201
	Part B	0.623
	Total	0.605
Cholecystectomy	Part A	0.229
	Part B	0.583
	Total	0.565
Hernia repair	Part A	0.384
	Part B	0.596
	Total	0.571

Total reliability was calculated according to the weighting of Part A (communication skills) to Part B (content of informed consent) set as 3:7.

3.2. Student performance

Students performed well in the VOSCE with total mean scores of 88.9% (±4.6%), individual results ranging from 76.6%–98.4% (Fig. 1A). There were no ceiling effects (right shift = core limitation at the top of a scale as indication of a relatively easy test) or floor effects (left shift = difficult examination). On comparison of the three checklists, the total mean scores were 91.0% (±4.0%), 88.4% (±4.4%), and 87.0% (±4.7%) for the appendectomy, cholecystectomy, and hernia repair checklist, respectively (Fig. 1B). One-way ANOVA demonstrated that total scoring was different in the three checklists ($P < 0.05$). Therefore, Scheffé's method was performed to compare the individual checklists with each other. The cholecystectomy and hernia repair checklists were similar in total scoring (mean absolute difference in percentage points 1.2 ± standard error 0.96; $P > 0.05$). However, the appendectomy checklist was apparently easier than the cholecystectomy (2.65 ± 0.94; $P < 0.05$) and also easier than the hernia repair checklist (3.8 ± 0.92; $P < 0.05$). Altogether, the mean absolute differences were very low (<3.8). The statistical results can possibly be attributed to the very high similarity of all checklists and only a relative small and almost negligible difference when compared with the hernia repair checklist. In fact, this result seems to be more a calculative effect owing to the homogenous distribution of data and therefore may not reflect any relevance to the assessment instrument. When referring to Figure 1B, which depicts student performance in the three checklists, it is obvious that the total mean scores and standard deviation were closely related and mostly overlapping.

3.3. Interrater agreement

Total scoring between both raters was similar among the three checklists. Mean total performance scores of the participants were 92.8 ± 4.4% and 89.2 ± 3.9% for the appendectomy, 89.1 ± 5.3% and 87.6 ± 4.0% for the cholecystectomy, and 87.9 ± 5.4% and 86.2 ± 4.5% for the hernia repair checklist (raters 1 and 2, respectively). The rating pattern of the two examiners is indicated in Figure 2. Although the median performance scores were similar, there was a tendency that the

Table 2 – Item difficulty (P) and item discrimination (r) for Part B (= content) of the checklists (Color version of Table is available online).

	Appendectomy[1]		Cholecystectomy[2]		Inguinal hernia repair[3]	
	P	r	*P*	r	*P*	r
Introductory explanations						
Diagnosis/indication	0 .79	.249	0.71	.184	0 .43	.297
Choice of procedure	0 .50	.209	0.35	.189	0 .46	.068
Conversion to open surgery[1,2]/ Laparotomy[3]	0 .96	.119	0.71	.253	0.31	.314
Extending the scope of surgery	0.45	.181	0.25	−.117	0.18	.250
General complications						
Positioning injury	0.27	.321	0.17	.038	0 .14	.125
Thrombosis/embolism	0.86	.238	0 .88	.275	0.73	.383
Haemorrhage	0.98	.082	0.98	.071	0.97	.068
Infection	0.98	.385	0.94	.391	0 .90	.432
Injury of vessels and nerves	0.93	.204	0.94	.035	1	.000
Scarring	0.84	.362	0.73	.298	0.65	.215
Adhesions/bowel obstruction	0.80	.336	0.44	.303	0.59	.080
Incisional hernia	0.86	.595	0.65	.291	0.69	.435
Specific complications						
Injury to neighbouring organs	0.45	.328	0.94	.130	0.96	.035
Injury to aorta or pelvic vessels[1]/ bile duct[2]/ nerves[3]	0.91	.325	0.85	.272	0.76	.251
Aerodermectasia[1,2]/ Intolerance of mesh[3]	0.82	.091	0.52	.277	0.80	.507
Shoulder pain[1,2]/ Chronic inguinal pain[3]	0.61	.122	0.44	.185	0.51	.268
Pneumothorax[1,2]/ Recurrence, dislocation of mesh[3]	0.43	.122	0.38	.483	0.69	−.113
Postoperative recommendations						
Return to normal diet / ambulation	0.70	.149	0.67	.104	0.59	.293
Postoperative impairment	0.52	.142	0.79	.157	0.14	.125

Color coding as follows:

$P \geq 0.8$	easy	$r \geq 0.3$	excellent
$0.4 \leq P < 0.8$	good	$0.2 \leq r < 0.3$	fair
$0.2 \leq P < 0.4$	hard	$0.1 \leq r < 0.2$	acceptable
$P < 0.2$	to be eliminated	$r < 0.1$	poor

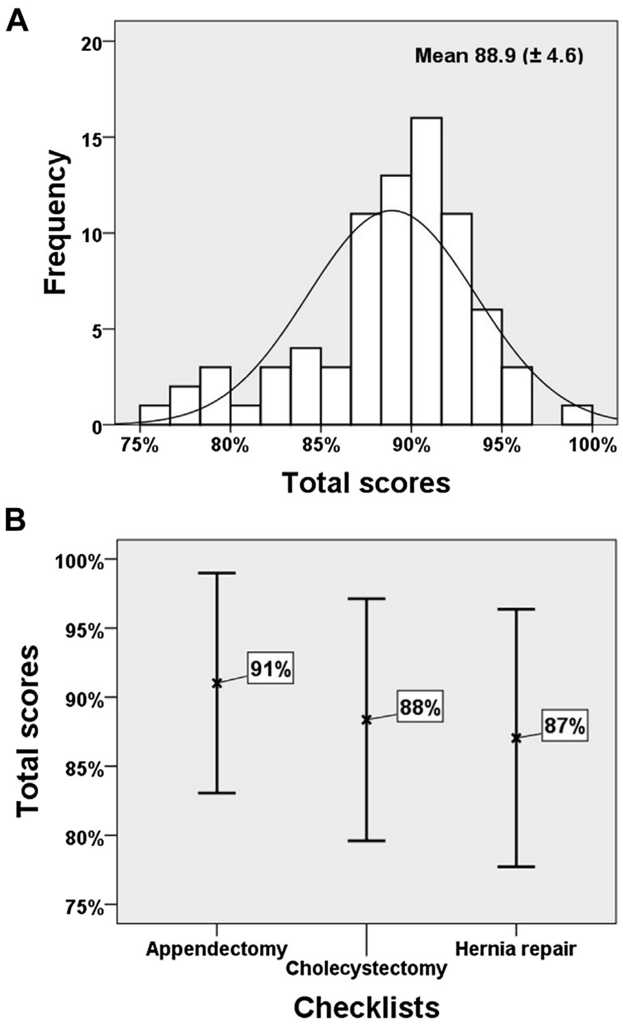

Fig. 1 – Distribution (frequency) of relative total scores as mean of both raters indicates a bell curve (A). Mean total scores ± 2 × standard deviation for the three checklists (B).

level of rating was slightly higher on all three checklists for rater 1 (surgical resident) when compared with rater 2 (dental resident). There was agreement in the utilization of the scale, with a greater dispersion on the hernia repair checklist.

Table 3 documents the correlation between both raters. The mean absolute differences (percentage points) were 3.63, 1.49, and 1.63 for the appendectomy, cholecystectomy, and hernia repair checklist, respectively. We found overall strong agreement on total scoring (ICC, 0.78). It is worth noting that Fleiss [23,24] considered an ICC value of >0.75 as excellent, between 0.4 and 0.75 as fair to good, and <0.4 as poor. Therefore, agreement was excellent for the appendectomy checklist (coefficients 0.83 and 0.84) and still on a very high level for the cholecystectomy (0.73 and 0.76) and hernia repair (0.76 and 0.77) checklists. PCC scores were also found to be very high with a total value of 0.80.

3.4. *Calculation of time to rate videos*

The videos were evaluated by the two raters during five self-determined sessions of paid overtime. In general, each rater assessed 8–10 videos in a row then took a short break lasting around 5–10 minutes.

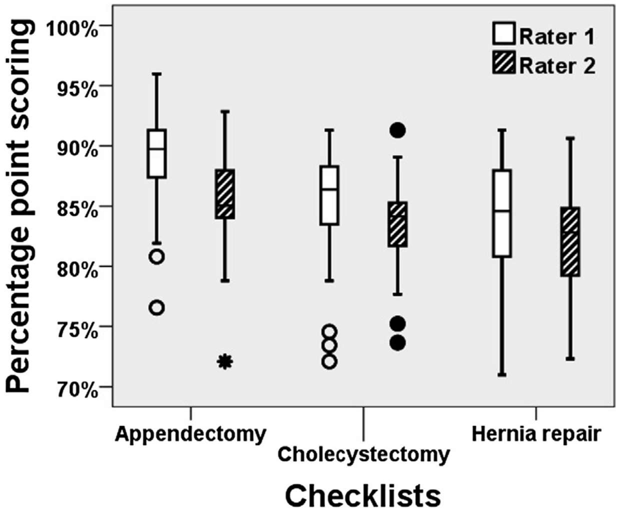

Fig. 2 – Box plot diagram of total scores in the three checklists to visualize rating patterns of the two raters. The bottom and top of the box mark the 25th and 75th percentile, respectively, and the middle dark line in the box indicates the median (50th percentile). The whiskers mark 1.5 times the interquartile range (IQR), circles mark outliers as 1.5–3 times, and stars mark outliers as >3 times IQR.

On examination, students interacted in pairs with SPs and had a total of 30 minutes to record two informed consent interviews, one of which the candidate pair then had to select for final assessment. Each of the 78 videos was approximately 10 minutes in length. Rating and marking on the checklists lasted on average 12 minutes per video. Thus, a total time of approximately 16 hours per rater was deemed necessary to assess the performance of the entire semester cohort.

4. Discussion

"Assessing the assessment" is vital, as the delivery of (V)OSCEs is complex and resource intensive. Any implementation or modification should be evaluated carefully to allow for quality assurance and check for feasibility in the local teaching environment. Thus, strategies may have to be developed on how best to implement standardization, which is known to affect the overall reliability of an OSCE positively. Well-designed checklists, video recording, and professional raters (as in trained and financially compensated) may enhance the overall quality of the OSCE.

4.1. *Establishment of a reliable and high-quality scoring system for raters*

In terms of the technical quality of the rating instruments, all three checklists demonstrated sound internal consistency with reasonably strong agreement on item difficulty and item discrimination as an indication of the high quality of all three checklists. In other terms, the checklists were of adequate

Table 3 – Interrater agreement of scoring for the three checklists.

Checklist	MAD	Range (95% CI)		ICC	Range (95% CI)		PCC	P
Appendectomy	3.63	2.98	4.28	0.83	0.66	0.92	0.84	<0.001
Cholecystectomy	1.49	0.47	2.50	0.73	0.46	0.87	0.76	<0.001
Hernia repair	1.63	0.65	2.60	0.76	0.54	0.89	0.77	<0.001
Total	2.31	1.79	2.83	0.78	0.67	0.85	0.80	<0.001

MAD = mean absolute difference; CI = confidence interval.

difficulty and differentiated well between respectable/good and weak student performance.

However, we were also able to identify a few checklist items with poor evaluation characteristics. We determined the following items to be of low difficulty ($P < 0.4$) combined with poor discrimination values ($r < 0.2$): “positioning injury” on the cholecystectomy and hernia repair checklists, “extending the scope of surgery” and “choice of procedure” on the cholecystectomy checklist, and “postoperative impairment” on the hernia repair checklist. The rationale behind this observation may be viewed from a teaching perspective. All these items are highly relevant in day-to-day surgical practice; however, it seems that practical experience is probably undervalued in the teaching context. This will have implications for teaching in the near future. After revisiting the curricular mapping of learning objectives, we will have to emphasize these points and ensure that students understand the operative concepts behind these procedures.

Total reliability of the VOSCE ranged from 0.565–0.605 for the individual checklists. These values were in line with recent literature. Following an evidence-based OSCE, which was also performed as a single station, Cronbach’s α was 0.58 and considered as acceptable [25]. In general, the evaluation of student performance is not based on single but on multiple assessment sessions, elements, or methods. In accordance, the assessment of clinical skills is commonly performed within the context of a multistation OSCE, and in this case, Cronbach’s α should at least overstep 0.6 [11] or better still reach at least 0.7 [25]. Therefore, we may extrapolate our results to a mini-OSCE round by combining all the three checklists. Thus, a theoretical total reliability of >0.9 could be expected and considered as very high (Spearman–Brown prediction formula = $r \times n/(1 + (n - 1) \times r)$ [26].

In our study, interrater agreement was excellent (ICC, 0.78; PCC, 0.80). In the literature, ICC scores ranging from 0.7 (OSCE assessing musculoskeletal ultrasound skills [27]) to 0.96 (evidence-based medicine OSCE [25]) have been reported. Two raters with totally different backgrounds (surgical versus dental) were used, yet we were still able to demonstrate strong agreement with one another. This we believe underlines not only the high quality of the checklists but also the overall design of the VOSCE including prior training of the raters. Involving albeit fewer and trained raters may prove to be a strategy to improve reliability considerably and thus increase the quality of an OSCE. Furthermore, we would like to emphasize that raters from a field other of surgery can still prove suited to the task of subject-specific skills and knowledge assessment.

Content validity was assured by an experienced surgeon, who developed the clinical scenarios, the roles of the SPs, and the checklists incorporating feedback from peer experts. However, student performance was not validated by comparing mean scoring of student cohorts from different educational levels (construct validity).

4.2. *Benefits of video recording and time-shifted rating of student performance*

The implementation of a VOSCE has considerable potential advantages for faculty members, educational coordinators, and candidates alike [28]. The classical OSCE carries a massive organizational burden associated with the necessity to guarantee and document the attendance of SPs and students. More importantly, the management of a large number of physician assessors with varying degrees of clinical and educational experience is a particular hurdle [11]. Not only must a predefined number of raters be in one place at one time despite concurrent clinical duties, ideally they should also have completed some skills training in assessment. In this context, implementing a VOSCE may be considered as an attractive alternative because the rating can occur outside of clinical normal working hours. However, shifting assessment duties into preferably financially compensated overtime has to be discussed carefully. In the context of the teaching module “Operative Medicine”, the time-shifted rating of student performance appeared to be the only solution to implement a practical clinical examination for the entire semester cohort. Moreover, it is generally accepted that such financial or time compensation can have a marked positive effect on the quality of the rating. Finally, this assessment method offers a substantial reduction in the administrative workload, as time-shifted rating does not necessarily require many raters.

From the perspective of raters, fatigue during the self-determined assessment sessions could also be reduced or even prevented [29]. The very nature of rating a video allows for breaks according to personal needs or preference. A positive consequence of this is a potential improvement in rating consistency [11]. From the student perspective, archived videos of their performance may be placed in an electronic portfolio [30]. Such a portfolio enables targeted feedback and may act as a personal guide throughout their degree course by highlighting knowledge gaps to both themselves and teaching staff.

The actual time required purely for time-shifted rating is not necessarily any shorter when compared with real-time live rating. However, in terms of logistics, time-shifted rating based on videos is a lot easier, as it is simply performed in series with no transition time or any additional time including disturbances. As an additional pilot within the context of this study, both raters reviewed 12 videos (four videos per checklist) chosen randomly 1 week after completion of the examination and re-rated them at a higher playback speed (1.2×). Preliminary data (not shown) demonstrated that the scores after re-rating were highly consistent with the primary scores, suggesting that the assessment time required could possibly be reduced. A direct and positive consequence of this would be a reduction in personnel costs involved [31,32]. Further studies using a crossover design will have to elucidate whether accelerated playback and/or other technical refinements may contribute further to greater time savings and improve rating convenience. Assessment of the videos with the help of electronic checklists enabling the fast selection of item and anchors or even an embedded digital rating tool using language recognition software will be evaluated in due course.

4.3. *Limitations*

Several limitations of our study should be mentioned. We did not investigate the students' attitude or viewpoint on video recording in this study. We may have to assume that student performance was in some way influenced by the process of being filmed. However, it is nowadays reasonable to accept that students on average are well aware of the benefit arising from the implementation of new technologies. Generally speaking, video recording is accepted as an approved tool to receive feedback, self-reflect on performance, and improve the accuracy of self-assessment [33,34]. Although the teaching module itself and subsequent summative examination is compulsory, consent to filming was not obligatory. Even so, the entire semester cohort participated in our study and appreciated the assurance of confidentiality and safe storage of the video material on providing their consent for the use thereof in our study.

Although all participating students were in the same semester (fifth year), we cannot exclude confounders such as differences in socioeconomic and educational background, as well as prior medical training or experience. As the entire semester cohort had to undergo examination, it was legally impossible to exclude students with previous training such as paramedic, nursing, or physiotherapy. However, this well reflects the genuine challenge that educators and examiners are confronted with: a heterogeneous population of students. It is perhaps worth restating at this point that we chose a random distribution of students to the three checklists to minimize effects from confounders.

Another limitation lies in the fact that the examination was a simulation as opposed to a real patient encounter. This limits the generalizability of our findings with respect to clinical context. Although students took the task of obtaining informed consent seriously, the scenario itself was still artificial. On review of the videos, it was noticeable that a number of students appeared to execute a memorized list of informed consent items as a monologue rather than a dialogue with the SP. It may well be the case that at least a proportion of candidates were influenced by the examination conditions to such an extent as to lose empathy with the SP.

Another limitation may lie in the fact that we only had two raters to assess the videos. Although we were able to demonstrate very strong interrater agreement, the generalizability of that finding remains limited. When recruiting, our aim was to select raters willing to assess the complete number of 78 videos during paid overtime. Following publication of the position advertisement, only two candidates qualified.

4.4. *Utility of the study*

The utility framework proposed by van der Vleuten [2] can be used to evaluate the value of assessment tools within a given curriculum. The formula to determine the utility (*U*) comprises five variables, those being reliability (*R*), validity (*V*), educational impact (*E*), acceptability (*A*), and cost (*C*). Reliability was reasonably high given the setting of a single-station VOSCE. Demonstration of validity was limited; however, the checklists were developed properly by an expert. Educational impact was high, as students were engaged in successful learning, as demonstrated by high average total performance scores. Acceptability was not a focus of the study. However, the VOSCE has been established as a routine summative assessment tool for the last 3 years. Cost-effectiveness of the VOSCE was a critical component as it is for any other form of practical clinical assessment. The raters received financial compensation (16 hours); however, this cost was effectively covered by reductions in administration costs. Moreover, implementation of high-stake OSCE stations into a multistation course may provide the opportunity to reduce the overall number of stations while retaining the high degree of internal consistency and decreasing costs.

The utility of the VOSCE is a multiplicative function of the above-mentioned variables. In our study, we demonstrated an approach to simplify the organization of an OSCE while guaranteeing high-quality measures for assessment. In doing so, we believe that the overall utility has not been jeopardized.

5. Conclusion

VOSCE is a feasible, objective, and reliable alternative to traditional live scoring in surgical education. In view of the German National Competency-based Catalogue of Learning Objectives [35], which will be published in due course, the development of standardized tools for the assessment of competencies and skills in surgery is becoming an essential element in the evaluation of undergraduate students [36].

Further research could explore to a greater extent the educational impact of VOSCE, for example, by investigating whether it generates stimuli to address the specific learning needs of the individual student. We also need to consider whether the narrative feedback given by SPs from their perspective may further enhance the acceptance of the VOSCE, as there are currently no means of direct interaction between raters and students possible.

Acknowledgment

The authors thank the students, student tutor Ann-Kathrin Kirchhoff, and in particular Susanne Koch as organizational assistant for the teaching module "Operative Medicine" for their continuing support in our research efforts to improve the feasibility and quality of the video-recorded objective structured clinical examination. They also express their gratitude for the financial support from the Office of the Dean of Studies (intramural funding of innovative teaching projects).

Author Contribution: All authors were involved in the form and/or study design and contributed critically to the final preparation of this article, including approving the final version of the manuscript. In particular, S.K. conceived and designed the study, wrote the final study protocol, ran the study, collected the results, analyzed the data, and drafted the manuscript. C.K. and A.S.-N. assisted in implementing the study. C.K., S.S., and Y.G. analyzed the data. T.R. advised at various stages, from protocol design to data interpretation, and assisted revising the manuscript. B.M.G. was as a general advisor most helpful in assisting throughout and in the interpretation of data. A.E. proofread the article and critically revised the correct use of terminology and description/discussion of data.

Disclosure

The authors of this manuscript have no conflicts of interest to disclose.

REFERENCES

[1] Hecker KG, Adams CL, Coe JB. Assessment of first-year veterinary students' communication skills using an objective structured clinical examination: the importance of context. J Vet Med Educ 2012;39:304.

[2] Van Der Vleuten CP. The assessment of professional competence: developments, research and practical implications. Adv Health Sci Educ Theor Pract 1996;1:41.

[3] Govindan VK. Enhancing communication skills using an OSCE and peer review. Med Educ 2008;42:535.

[4] Yudkowsky R, Alseidi A, Cintron J. Beyond fulfilling the core competencies: an objective structured clinical examination to assess communication and interpersonal skills in a surgical residency. Curr Surg 2004;61:499.

[5] Chipman JG, Beilman GJ, Schmitz CC, Seatter SC. Development and pilot testing of an OSCE for difficult conversations in surgical intensive care. J Surg Educ 2007;64:79.

[6] Tchorz KM, Binder SB, White MT, et al. Palliative and end-of-life care training during the surgical clerkship. J Surg Res 2013;185:97.

[7] Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65:S63.

[8] Shah B, Miler R, Poles M, et al. Informed consent in the older adult: OSCEs for assessing fellows' ACGME and geriatric gastroenterology competencies. Am J Gastroenterol 2011;106:1575.

[9] 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.

[10] Pell G, Fuller R, Homer M, Roberts T. How to measure the quality of the OSCE: a review of metrics—AMEE guide no. 49. Med Teach 2010;32:802.

[11] Vivekananda-Schmidt P, Lewis M, Coady D, et al. Exploring the use of videotaped objective structured clinical examination in the assessment of joint examination skills of medical students. Arthritis Rheum 2007;57:869.

[12] Humphris GM, Kaney S. The Objective Structured Video Exam for assessment of communication skills. Med Educ 2000;34:939.

[13] Karabilgin OS, Vatansever K, Caliskan SA, Durak HI. Assessing medical student competency in communication in the pre-clinical phase: objective structured video exam and SP exam. Patient Educ Couns 2012;87:293.

[14] Goettingen Catalogue of Learning Objectives. Available at: http://www.med.uni-goettingen.de/de/media/G1-2_lehre/lernzielkatalog.pdf. Accessed: December, 2013.

[15] Classical test theory—Wikipedia the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Classical_test_theory#Evaluating_items:_P_and_item-total_correlations. Accessed: January, 2014.

[16] Item Discrimination Indices, Institute for objective measurement, Inc. USA - Available at: http://www.rasch.org/rmt/rmt163a.htm. Accessed: January, 2014.

[17] Survery methods. What is Cronbach's alpha - Available at: http://surveymethodsaddicts.blogspot.co.uk/. Accessed: January, 2014.

[18] What does Cronbach's alpha mean? UCLA - Available at: http://www.ats.ucla.edu/stat/spss/faq/alpha.html. Accessed: January, 2014.

[19] Intraclass correlation coefficiency—Wikipedia the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Intraclass_correlation. Accessed: January, 2014.

[20] Pearson product-moment correlation coefficient—Wikipedia the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient. Accessed: January, 2014.

[21] Analysis of variance—Wikipedia the free encyclopedia. Available at: http://en.wikipedia.org/wiki/ANOVA. Accessed: January, 2014.

[22] Scheffé's method—Wikipedia the free encyclopedia. Available at: http://en.wikipedia.org/wiki/Scheff%C3%A9%27s_method. Accessed: January, 2014.

[23] Fleiss JL. The design and analysis of clinical experiments. New York: John Wiley & Sons; 1986.

[24] Fleiss JL. Measuring agreement between two judges on the presence or absence of a trait. Biometrics 1975;31:651.

[25] Tudiver F, Rose D, Banks B, Pfortmiller D. Reliability and validity testing of an evidence-based medicine OSCE station. Fam Med 2009;41:89.

[26] Moeltner A, Schellberg D, Juenger J. Basic quantitative analyses of medical examination. GMS Z Med Ausbild 2006;23:Doc53.

[27] Kissin EY, Grayson PC, Cannella AC, et al. Musculoskeletal ultrasound objective structured clinical examination: an assessment of the test. Arthritis Care Res (Hoboken) 2014;66:2.

[28] Casey PM, Goepfert AR, Espey EL, et al. To the point: reviews in medical education—the Objective Structured Clinical Examination. Am J Obstet Gynecol 2009;200:25.

[29] McLaughlin K, Ainslie M, Coderre S, Wright B, Violato C. The effect of differential rater function over time (DRIFT) on objective structured clinical examination ratings. Med Educ 2009;43:989.

[30] Sanchez Gomez S, Ostos EM, Solano JM, Salado TF. An electronic portfolio for quantitative assessment of surgical skills in undergraduate medical education. BMC Med Educ 2013;13:65.

[31] Rau T, Fegert J, Liebhardt H. How high are the personnel costs for OSCE? A financial report on management aspects. GMS Z Med Ausbild 2011;28:Doc13.

[32] Kelly M, Murphy A. An evaluation of the cost of designing, delivering and assessing an undergraduate communication skills module. Med Teach 2004;26:610.

[33] Maloney S, Paynter S, Storr M, Morgan P. Implementing student self-video of performance. Clin Teach 2013;10:323.

[34] Hawkins SC, Osborne A, Schofield SJ, Pournaras DJ, Chester JF. Improving the accuracy of self-assessment of practical clinical skills using video feedback—the importance of including benchmarks. Med Teach 2012;34:279.

[35] German National Competency-based Catalogue of Learning Objectives. Available at: http://www.nklm.de/. Accessed: December, 2013.

[36] Kadmon M, Bender MJ, Adili F, et al. [Competency-based medical education: National Catalogue of Learning Objectives in surgery]. Chirurg 2013;84:277.