An Objective Structured Clinical Exam to Assess Semiology Skills of Medical Students

Um Exame Clínico Objetivo Estruturado para Avaliar Aptidões de Semiologia em Alunos de Medicina

ARTIGO ORIGINAL

Vitor Hugo PEREIRA[1,2], Pedro MORGADO[1,2], Mónica GONÇALVES[1,2], Liliana COSTA[1,2], Nuno SOUSA[1,2], João CERQUEIRA✉[1,2]

Acta Med Port 2016 Dec;29(12):819-825 ▪ http://dx.doi.org/10.20344/amp.8407

ABSTRACT

Introduction: Mastery of history taking and physical exam skills is a key competence of medical students. Objective Structured Clinical Examinations are the gold standard to assess these competencies, but their implementation in Portugal is poorly documented. We describe the implementation and our seven years experience with a high-stakes Objective Structured Clinical Examination to assess these skills in the School of Medicine, University of Minho.

Material and Methods: Our Objective Structured Clinical Examination is in place since 2010 and has been subject to continuous improvements, including the adoption of a standard setting procedure and an increase in the number of stations.

Results: Grades in our exam are well distributed and discriminate among students. History taking grades are lower and have remained stable throughout the years while physical examination scores have risen. The exam is reliable, with internal consistency above 0.45 and a G-coefficient of 0.74. It is also feasible, with a total testing time of approximately 20 hours for 140 students, and the involvement of 18 standardized patients and 18 faculty assessors. More importantly, it was able to engage the students, who recognize its importance.

Discussion: The most important validity criterion of our, and any Objective Structured Clinical Examination, would be predictive validity, the ability to predict the performance of students in the clinical context.

Conclusion: Our approach to a high-stakes Objective Structured Clinical Examination shows that it is feasible, reliable, valid and fair and can be implemented with success in the Portuguese setting.

Keywords: Education, Medical, Undergraduate; Educational Measurement; Medical History Taking; Physical Examination; Portugal; Reproducibility of Results.

RESUMO

Introdução: O domínio das aptidões de colheita de história e do exame físico é uma competência chave para os estudantes de medicina. Os Exames Clínicos Objetivos Estruturados são o padrão para avaliar estas competências, mas a sua implementação em Portugal está pouco documentada. Descrevemos a implementação e a nossa experiência de sete anos, com um Exame Clínico Objetivo Estruturado para avaliar estas aptidões na Escola de Medicina da Universidade do Minho.

Material e Métodos: O nosso Exame Clínico Objetivo Estruturado existe desde 2010 e tem sido sujeito a melhorias contínuas, das quais se destacam a adoção de um procedimento de *standard setting* e o aumento do número de estações.

Resultados: As notas no nosso exame estão bem distribuídas e discriminam entre estudantes. As notas da colheita da história são inferiores e têm permanecido estáveis ao longo do tempo, enquanto as do exame físico têm aumentado. O exame é fiável, tendo uma consistência interna acima de 0,45 e um coeficiente G de 0,74. Também é praticável, tendo um tempo total de teste de 20 horas para 140 alunos, envolvendo 18 pacientes estandardizados e 18 examinadores. Mais importante, o Exame Clínico Objetivo Estruturado foi capaz de cativar os estudantes, que reconhecem a sua importância para a sua formação.

Discussão: O mais importante critério de valor do nosso, e de qualquer outro Exame Clínico Objetivo Estruturado, será o seu valor preditivo, que se traduz na faculdade de prever o desempenho dos estudantes em contexto clínico.

Conclusão: A nossa abordagem para um Exame Clínico Objetivo Estruturado mostra que é praticável, fiável, válido e justo e que pode ser implementado com sucesso no contexto português.

Palavras-chave: Anamnese; Avaliação Educacional; Educação Médica Pré-Graduada; Exame Clínico; Portugal; Reprodutibilidade de Resultados.

INTRODUCTION

The ability to explore the complaints of a patient, and to elicit and interpret the relevant symptoms and signs by means of a thorough clinical history and physical examination is, even in an era of medical technology, the mainstay of medical diagnosis.[1] Mastering of clinical history taking and physical examination is therefore a critical competency for medical students and has to be extensively trained and adequately assessed throughout the medical degree.

Objective Structured Clinical Examinations (OSCEs) are the gold standard for assessment of clinical competency, particularly in semiology,[2] but implementation of a high-stakes OSCE is highly resource-consuming and requires a strong commitment of the medical school. Perhaps because of these difficulties, OSCE implementation in Portuguese Medical Schools has lagged behind and, to the best of our knowledge, there are no published reports on the feasibility of a high-stakes OSCE in the national setting. In 2010,

1. School of Medicine and Life and Health Sciences Research Institute (ICVS). University of Minho. Braga. Portugal.
2. ICVS/3B's Associate Laboratory. Braga/Guimarães. Portugal.
✉ Autor correspondente: João Cerqueira. jcerqueira@med.uminho.pt
Recebido: 01 de novembro de 2016 - Aceite: 30 de novembro de 2016 | Copyright © Ordem dos Médicos 2016

MedUM implemented a high-stakes semiology OSCE, backed by a solid (SP) program, and specifically designed to assess the mastery of history-taking and physical exam skills by 3rd year medical students, just before the start of their clinical rotations. The present paper describes the development and implementation of the MedUM semiology OSCE and its seven-year experience. Our aim is to document the feasibility of such endeavor in the Portuguese setting and, putting an emphasis on the successive improvements to the process and lessons learned, help other schools to implement similar programs, avoiding some of our pitfalls.

MATERIAL AND METHODS

OSCE organization: current status and historical perspective

MedUM medical degree is a six-year competency-based program. As previously described, OSCE are one of the best methods to evaluate clinical competencies in healthcare professionals. As so, we introduced an OSCE at the end of the curricular unit (CU) Introduction to Clinical Medicine. This CU takes place in the 3rd year of the curriculum, before students start clinical rotations in the hospital. The learning goals/competencies of this CU are the following:

1. Take a complete clinical history in different settings;
2. Explore and thoroughly characterize the patients' complaints;
3. Perform a systematic head-to-toe physical examination;
4. Integrate symptoms and signs into syndromes and diseases;
5. Apply techniques of clinical communication in the context of a patient visit;
6. Summarize and concisely report clinical information.

Currently, our OSCE (designed to assess these learning objectives) is a six-station exam composed by three similar versions (i.e. replicates) occurring simultaneously. This means that at the same time, 18 SPs and 18 faculties are rating the same number of students. Each station has the total duration of 15 minutes and is divided in two different tasks. The first consists in taking a clinical history directed to the chief-complaint of the SP (10 minutes); the second task consists in performing a specific maneuver or a group of maneuvers from our head-to-toe physical examination script (five minutes). Although assessed in the same room and with the same SP/assessor pair, the clinical history taking and the physical examination tasks are unrelated. After the end of the station the students have five minutes to answer a post-encounter challenge (only present in three out of six stations, in order to decrease the work burden and allow for some rest during the exam). Post-encounters are short-answer questions to assess the ability to summarize data, where students are asked to describe, in a few lines, either: i) the chief-complaint, ii) the past clinical history or iii) the findings of the physical examination. Student performance is graded based on history taking (addressing learning goals 1, 2, 4, 5), physical examination (learning goals 3, 5), communication (learning goal 5) and data synthesis skills (learning goals 4, 6). The final score of the exam is a 20 values scale, where the pass/fail cut-off is set in 9.5 values. Contrary to other exams along the curricula, if a student fails the exam he or she will not be allowed to progress further to the 4th year, making this OSCE a high-stakes exam.

Along the years the OSCE suffered some amendments until it reached the current format (Table 1). We designed the exam in the year 2009/2010 and at the time it had 5 stations (10 min + 5 min) running in duplicate (i.e. two replicates). In the next year we introduced the post-encounter tasks and the standard setting procedure, described in the results below. This last point came from the general impression that without the standard setting the exam lacked the discriminative power to pass/fail students and to rank them properly in our grading scale. In 2011/2012 we had to accommodate the increase in the number of students/year and we started to run the exam in three replicates. The number of SPs and faculty was increased and an effort was directed to improve the quality of SP training to assure reliability among the different versions of the same station. After successfully overcoming this challenge we felt confident to move forward and in 2012/2013, based on a G-theory analysis of previous OSCEs, we introduced a 6th station to further increase the reliability of the exam (see below). Two years later we designed and validated our own communication scale[3], which is now in use.

Case development and checklist validation

Case development and checklist validation is a step-wise process that involves the faculty and the SPs. The first step is to select the cases according with the blueprint of the exam, which is based in two dimensions: age of the 'patient' and medical specialty of the scenario. The cases follow a standard script and are developed around a chief complaint and not a diagnosis. This has the advantage of avoiding the development of typical presentations of a disease (which narrows the differential diagnosis), promoting the

Table 1 – Changes introduced in the exam structure

2009/2010	5 stations OSCE with 2 replicates
2010/2011	Introduction of the post-encounter tasks and the standard setting method
2011/2012	Introduction of another replicate (3)
2012/2013	Introduction of a 6th station (to increase reliability)
2013/2014	No major change
2014/2015	Introduction of our communication assessment scale
2015/2016	No major change

creation of more non-specific cases with broader differential diagnosis. After the first draft, both the case and the checklist are reviewed by a different group of faculty to check the coherence of the script and the suitability of the checklist. Next, the case moves to the faculty responsible for training the SPs, who has the freedom to introduce slight changes in the case to increase its veracity and consistency. SPs used in the exam are laywomen and men, aged 18 years old and plus, without any major medical condition and not healthcare professionals, selected among those participating in our SP program for at least one year; as part of the program they receive an hourly compensation for participating in trainings, learning activities and exams. Specifically for the exam, training consists of at least three two-hour sessions with all SPs of a given history and one trainer/faculty. These can be complemented with additional sessions when needed. When the training is complete a new phase of the process ensues. To assess if the SPs playing the same case are properly standardized we perform a default interview (very similar to a questionnaire) and compare their performance. Then we perform a free interview with a faculty that does not know the case to assess the performance of the SP in a more realistic and unpredictable scenario and to validate our checklist. This interview is reviewed and the items of the checklist rated one by one according with their appropriateness for the exam. At this time some items are eliminated, others are rephrased and others added to the final version. After all this process the SPs receive feedback about possible errors detected and additional training sessions are scheduled if needed. The faculty is debriefed previously to the exam to clarify any doubts about the proper filling of the checklist.

Data analysis and statistics

In this paper we will provide and analyze some data collected during the last years, according to published recommendations for publications about OSCEs.[4] To facilitate comparisons, all scores were converted into fraction of maximal points possible and presented in a 0-1 scale. Values are mean and, when applicable, standard deviations. Internal consistency was analyzed with Cronbach's alpha, while G-theory was used to calculate the G-coefficient and select an adequate number of stations. Psychometricians developed G-theory as a complement to item-response theory. Its objective is to analyze the different sources of variability in the scores of a particular exam and extract the proportion that is attributable to real differences in students' performance (the G-coefficient). The advantage of G-theory, and the reason for its widespread use in evaluating OSCEs, is its flexibility: the models underlying the analysis can be adapted to a multitude of designs. In addition, G-theory results can be extrapolated to different scenarios, allowing predictions regarding the impact of alterations in the number of stations or examiners. In our case, we used a nested design with students*station type (nested on stations) and extrapolated for different numbers of stations.

All statistical analyses were performed using IBM SPSS statistics software, v22.0 except G-theory analyses (based on the sum of squares obtained from the IBM SPSS statistics software) which were run on a Excel sheet template courtesy of Dr. David Swanton, formerly at the National Board of Medical Examiners. Students' opinions are recorded in written form at the end of the exam and are voluntary, being qualitatively analyzed afterwards. Representative opinions are transcribed here.

RESULTS

Scoring

Each component of the exam contributes to the final grade, with different assigned weights: history-taking abilities contribute 55%, physical exam performance contributes 25%, communication skills contribute 15% (with equal weights for the score given by the faculty and by the SP) and post-encounter tasks contribute 5%.

Since 2011 a standard setting procedure, according to the borderline methodology[5] is individually applied to each of the history taking and physical examination checklist scores before computing the final grade. The remaining exam components are not subject to any standard setting procedure. In brief, the borderline standard setting procedure consists of three phases. In the first, during the exam, the performance of each student in each task is classified as clear fail, borderline or clear pass (independently and in addition to grading the task-specific checklist), after previous training and calibration of the examiners. In the second, the average checklist score of all borderline students in a given task is defined as the standard for that task (cut-off). Finally, all students, independently of their classification as clear fail, borderline or clear pass, whose scores in a given task are lower than the standard for that task, are given a score 0 and all the others are given a classification that increases linearly from 0.5 to 1 and is proportional to the amount of the checklist score that is higher than the standard. As can be gleaned from Fig. 1, introduction of this procedure had no impact in the average grades of the two components, nor on the overall score.

Overall student performance has remained relatively stable throughout the years (Fig. 1), with a peak in students' grades in 2012; most likely, this is due to a cohort effect, as no changes were introduced in the exam structure that year and scores returned to lower values in the following years. More importantly, final exam grades have been consistently well distributed and adequately spread across the full spectrum of possible grades (Table 2). Analysis of student performance on the different domains (Fig. 1) revealed that history-taking skills have consistently scored lower that all the other domains and remained stable throughout the years. On the contrary, physical examination skills have seen a rise in scores after the two first years and a stability ever since, with a similar trend for post-encounter scores. Interestingly, communication skills scores have seen a descendent trajectory, to become aligned with the physical examination and post-encounter scores, and the

disappearance of the gap between scores given by the faculty and SP. Significantly nor the introduction of the standard setting procedure, nor that of a 6th station seem to have had a significant impact in students scores.

Reliability

Internal consistency of the exam was adequate in all cohorts and high in most of them. The lowest value for Cronbach's alpha was obtained in 2011, the year after the introduction of the standard setting procedure. In 2013 we performed a G-theory analysis of 2012 exam scores, based on a nested design, and determined the G-coefficient to be

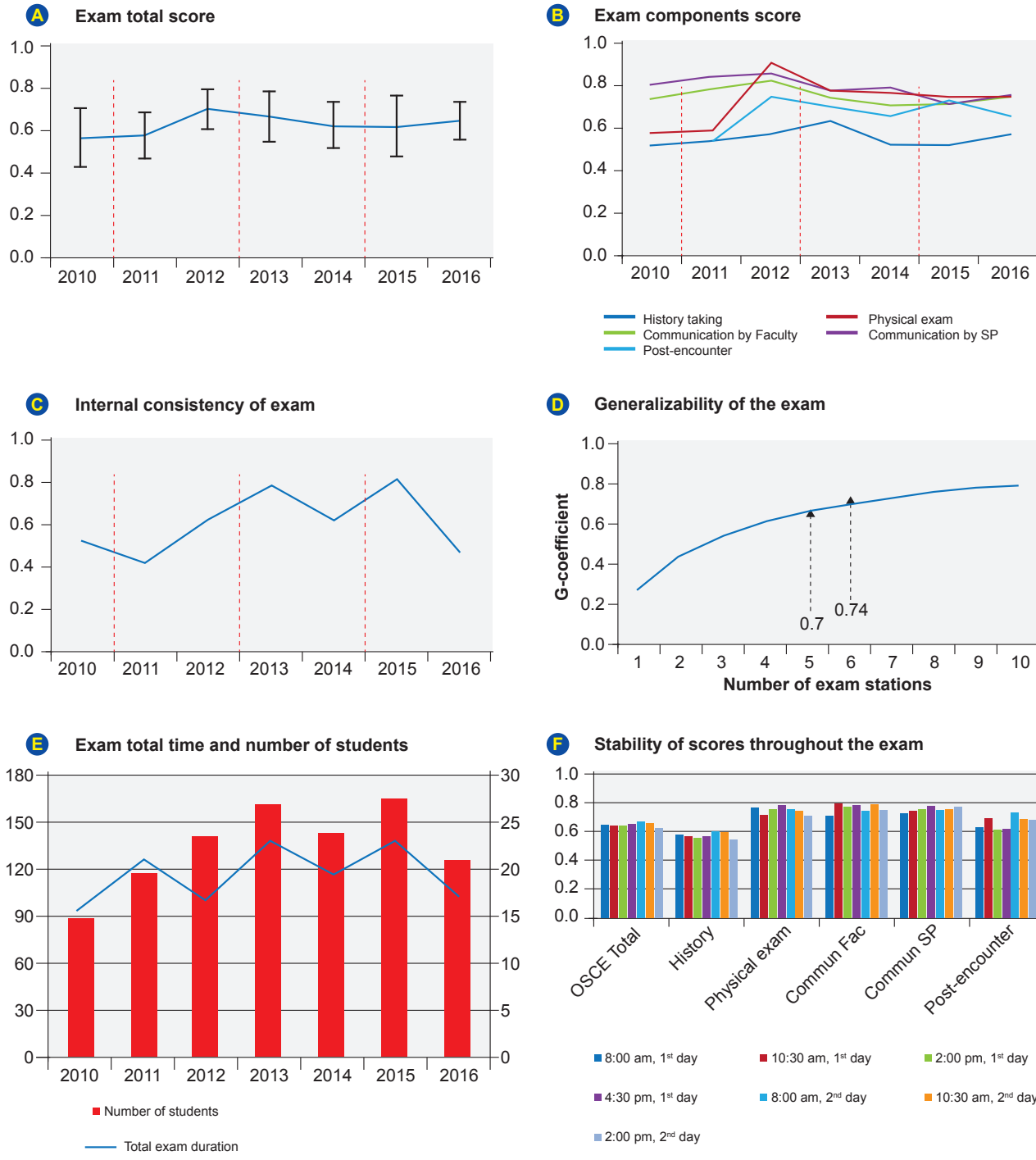

Figure 1 – Summary of results. Evolution of the total OSCE scores (A) and each individual component score (B) since 2010. Evolution of the internal consistency of the exam since 2010 (C). Plot of the G-coefficient of the exam according to the number of stations (D). Evolution of the number of students and testing hours since 2010 (E) Stability of the scores throughout the 7 shifts in the 2016 exam, for the overall scores and each individual component score (F). Dashed lines represent changes in the structure of the OSCE. Error bars are standard deviations.

Table 2 – Student grades

Year of the exam		History taking	Physical exam	Communication by Faculty	Communication by SP	Post-encounter	TOTAL
2010	Mean	0.528	0.5762	0.7406	0.8051		0.5669
	Std. Deviation	0.16915	0.19697	0.11916	0.11642		0.1347
	Minimum	0	0	0	0		0
	Maximum	0.9	0.96	0.89	0.95		0.81
2011	Mean	0.5412	0.5878	0.7881	0.8461	0.5363	0.5799
	Std. Deviation	0.15617	0.18991	0.0779	0.08811	0.13187	0.10978
	Minimum	0.11	0	0.51	0.48	0.19	0.17
	Maximum	0.8	0.93	0.93	1	0.83	0.8
2012	Mean	0.5752	0.9087	0.8227	0.86	0.7489	0.7025
	Std. Deviation	0.15272	0.06544	0.08516	0.09081	0.14296	0.09672
	Minimum	0.1	0.52	0.55	0.41	0.07	0.38
	Maximum	0.86	1	0.96	0.99	1	0.89
2013	Mean	0.6347	0.783	0.7436	0.7776	0.7043	0.668
	Std. Deviation	0.109	0.21528	0.13474	0.13586	0.13307	0.11893
	Minimum	0.11	0.1	0.3	0.32	0.27	0.26
	Maximum	0.87	1	1	1	1	0.85
2014	Mean	0.527	0.7634	0.7118	0.797	0.6655	0.6227
	Std. Deviation	0.14949	0.14734	0.13234	0.10774	0.1967	0.1077
	Minimum	0	0.25	0	0	0	0.18
	Maximum	0.79	1	0.9	0.95	1	0.85
2015	Mean	0.5253	0.7497	0.7154	0.7156	0.7354	0.6225
	Std. Deviation	0.16417	0.19303	0.16102	0.16037	0.20013	0.14757
	Minimum	0	0	0	0	0	0
	Maximum	0.75	1	0.9	0.9	1	0.81
2016	Mean	0.5682	0.7428	0.759	0.757	0.6618	0.6441
	Std. Deviation	0.12257	0.14766	0.14542	0.06455	0.16848	0.08917
	Minimum	0.18	0.35	0.2	0.55	0	0.33
	Maximum	0.8	0.97	0.97	0.93	0.93	0.84

0.7, at the lower limit of the adequate range (Fig. 1). This analysis also revealed that increasing a single station (from 5 to 6) was sufficient to increase the generalizability of the exam to values considered more adequate, which prompted us to introduce this 6th station right on that year.

Validity

The exam committee elaborates the blueprint of the exam. Communication skills assessment and post-encounter tasks are fixed components of the exam and are not subject to blueprinting. The history taking component has a two dimensional blueprint in which the chief complaint to be explored during the encounter is classified in one of ten categories (mostly according to the systems of the body) and the age of the patient is classified in 10 year intervals (see Table 3, for an example). Clinical scenarios for this component are custom built for every exam according to this blueprint, trying to cover most ages and six different complaint categories. The physical exam component has also a two dimensional blueprint in which the maneuvers to be performed are categorized in one of five categories (mostly according to the parts of the body) and whether the patient has normal or any abnormal findings. Physical exam tasks are selected to cover each of the five categories with two stations covering either limbs/neurological or thorax (see Table 4 for an example). Of note, in every exam there is always a station to measure the vital signs of the patient.

In the same CU, students are also assessed in a computerized test with 100 multiple choice questions about semiology and clinical reasoning (knowledge) and have a professionalism global rating score given by their tutors in the (small group) clinical skills lab sessions, during which history taking and physical exam are trained (professionalism). In order to provide more insight into the validity of the exam, we correlated these 'external' scores with the total OSCE scores and the scores of each component, taking the 2016 exam as an example (Table 5). Overall OSCE scores were significantly correlated with both professionalism and knowledge scores, but more strongly with the former. More interesting were the correlations with

Table 3 – History taking blueprint

	20 – 29	30 – 39	40 – 49	50 – 59	> 60
Musculoskeletal					X
Nervous/Special senses	X				
Psychiatric					
Skin		X			
Endocrine					
Digestive					
Urinary				X	
Reproductive					
Cardiovascular			X		
Respiratory					
General complaints			X		

the individual OSCE components: while the post-encounter task was significantly correlated with knowledge, but not professionalism, history taking and communication skills were significantly correlated with professionalism but not with knowledge.

Feasibility

Since its implementation in 2010, a total of 933 students have been assessed in our OSCEs. Despite the increase in student numbers, the introduction of a third replicate series in 2012 has allowed total testing time to remain relatively stable, between 20 and 22 hours, even after the introduction of a 6th station in 2013 (Fig. 1). Given the need to distribute the exam throughout several days and in order to increase fairness perception, all students receive the six history vignettes 48 hours in advance of the first day of testing, although the history taking and physical exam rating scales are never made public. In order to document that this methodology does not introduce any scoring bias, we analyzed average OSCE total scores in the 7 shifts that comprise the 2016 OSCE, without any significant difference among them (Fig. 1).

Student opinions

Students are very positive about the exam and feel that the exam is well organized. They are also very aligned with the exam aims and have the sense that 'the exam assesses exactly what is most important for our future profession'. Despite all the stress, they feel like being in the 'real setting'. Some think the exam time is enough, while for others it is long and tiring. Finally, most students value the training sessions in the clinical skills lab, which help them prepare for the exam.

Table 4 – Physical exam blueprint

	Normal	Abnormal
Vital signs	X	
Head and neck	X	
Thorax		X
Abdomen	X	
Limbs/neurological	X X	

DISCUSSION

In the present paper we describe the implementation and development of a high-stakes OSCE to assess semiology skills of undergraduate medical students in Portugal. Besides presenting our OSCE, we also detail how it grew both in number of students assessed and in stations and the rationale behind each change. We also analyze the impact of such changes in the scores and the reliability of the exam. In this regard, one of the most interesting findings is the apparent small impact of the standard setting procedure. Despite this, however, we decided to keep it, as we had the perception that it increased fairness and discrimination among students, which is in accordance with the published literature.[5]

A longitudinal analysis of student grades reveals a relative stability of the overall classifications but interesting trends in the individual components of the exam. While history-taking scores have remained stable, physical exam performance has steadily increased. This can be attributed to the fact that students have a better knowledge of the limited repertoire of possible physical examination questions, but also to a positive impact of the OSCE on student motivation to learn and master the physical examination. On the contrary, communication skills scores have slightly declined throughout the years, probably reflecting the increased experience and confidence of both SPs and Faculty in assessing such difficult subjects. Interestingly, this evolution was also paralleled by a convergence between the communication scores given by the SP and Faculty.

Table 5 – Correlation between OSCE scores and knowledge and professionalism grades (correlation coefficient for statistically significant correlations, $p < 0.05$)

	Knowledge	Professionalism
History taking	NS	0.232
Physical examination	NS	NS
Communication by Faculty	NS	0.285
Coomunication by SP	NS	0.311
Post-encounter	0.286	NS
Total OSCE score	0.260	0.324

In this case, besides an increased experience, the use, in the two last years, of a new, custom-built communication assessment scale,[3] which is shorter and more easy to use, might also have contributed to this finding. In light of these results, it is important to highlight the need to extensively train the SPs and Faculty, not only in depicting the clinical scenario, but also in the proper use of the assessment instruments including checklists and scales.

The most important validity criterion of our, and any OSCE, would be predictive validity, the ability to predict the performance of students in the clinical context. In the absence of such information, we analyzed the correlation between OSCE scores and the grades of the same students in a knowledge test about semiology and a professionalism score during clinical training. Our results support the validity of the exam. Indeed, while the post-encounter task (the only that is performed without facing the SP) was the only component significantly correlated with the knowledge grade, history taking and communication skills were significantly correlated with professionalism, but not knowledge. In addition, our OSCE has strong content validity as it covers a large set of possible complaints and patient ages, as well as different physical examination tasks.

Finally, we show that, in our context, these exams are feasible in a reasonable amount of time despite the large number of students, faculty and SPs involved. More importantly, we also show that, by being transparent and documenting all processes, it is possible to align students with the aims of the exam and count with their favorable opinion. In our idea, this is the only way to make such a big enterprise last for so many years and to collect all its benefits, both for the individual student and the medical school community as a whole.

CONCLUSION

In the Portuguese setting it is possible to implement a reliable, valid and fair high-stakes OSCEs to assess medical semiology and to engage the students into it.

PROTECTION OF HUMANS AND ANIMALS

The authors declare that the procedures were followed according to the regulations established by the Clinical Research and Ethics Committee and to the Helsinki Declaration of the World Medical Association.

DATA CONFIDENTIALITY

The authors declare having followed the protocols in use at their working center regarding patients' data publication.

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

FUNDING SOURCES

No subsidies or grants contributed to this work.

REFERENCES

1. Paley L, Zornitzki T, Cohen J, Friedman J, Kozak N, Schattner A. Utility of clinical examination in the diagnosis of emergency department patients admitted to the department of medicine of an academic hospital. Arch Intern Med. 2011;171:1394-6.
2. Harden RM. Misconceptions and the OSCE. Med Teach. 2015;:1-3.
3. Gonçalves M, Gonçalves M, Sousa AL, Morgado P, Costa P, Cerqueira JJ. Development and validation of a new instrument to assess communication skills. Madrid: XXI Congreso de la Sociedad Española de Educación Médica. 2013.
4. Patricio M, Juliao M, Fareleira F, Young M, Norman G, Vaz Carneiro A. A comprehensive checklist for reporting the use of OSCEs. Med Teach. 2009;31:112-24.
5. Kaufman DM, Mann KV, Muijtjens AM, van der Vleuten CP. A comparison of standard-setting procedures for an OSCE in undergraduate medical education. Acad Med. 2000;75:267-71.