European Journal of Ophthalmology / Vol. 16 no. 4, 2006 / pp. 595-603

A pilot study of the use of objective structural clinical examinations for the assessment of ophthalmology education

P. AYDIN[1], I. GUNALP[2*], B. HASANREISOGLU[3*], M. UNAL[3*], M. EROL TURACLI[2*],*
**Members of the Turkish Board of Ophthalmology Executive and Assessment Committee*

[1]Eye Department, Mesa Hospital, Ankara
[2]Eye Department, Ankara University Medical School, Ankara
[3]Eye Department, Gazi University Medical School, Ankara - Turkey

PURPOSE. *A pilot study was carried out to evaluate the practicality, reliability, and validity of an objective structured clinical examination (OSCE) for assessing the clinical skills and abilities of specialists in ophthalmology.*
METHODS. *Ten unfolded OSCE style, criterion referenced questions were asked to nine candidates to assess their clinical skills and abilities, as opposed to subject knowledge. Candidate and assessor reactions to the examination process were monitored and analyzed using participant observation and questionnaires administered immediately after the event. Relevant statistical techniques were applied to the results.*
RESULTS. *A total of 89% of candidates passed the examination, with the pass boundary set at 70%. Candidates revealed themselves more successful in meeting clinical skill criteria (mean 77%) than clinical ability criteria (mean 72%). Candidates, assessors, and observers all expressed the view that the OSCE pilot had been a successful way of assessing clinical skills and abilities.*
CONCLUSIONS. *OSCE style assessment is an effective and efficient means of assessing skills and abilities in clinical ophthalmology education. (Eur J Ophthalmol 2006; 16: 595-603)*

KEY WORDS. *Clinical assessment, Objective structured clinical examination, Ophthalmology education, OSCE, Postgraduate medical education*

Accepted: February 28, 2006

INTRODUCTION

With the growing number of postgraduate residents and the concomitant need to standardize the assessment of clinical skill and clinical behavior, the development of an objective clinical examination system has become essential for many countries. The objective structured clinical examination (OSCE) is a well-established form of assessment for undergraduates, which was specifically designed to evaluate clinical skills validly (1-7). However, although well established for undergraduate assessment, its value as the gold standard model for postgraduate, surgically oriented specialties such as ophthalmology remains to be explored and developed.

A pilot study was created to evaluate the practicality, reliability, and validity of an OSCE type of assessment for specialists in ophthalmology. The study covers two distinct aspects of the examination process. First, a description is afforded of the operating conditions and procedures used to establish a reliable and valid framework for the assessment. Secondly, an analysis of examinee and assessor reactions to the examination process is offered based on participant observation, and responses to questionnaires administered immediately after the event. By specifying the

 1120-6721/595-09$15.00/0

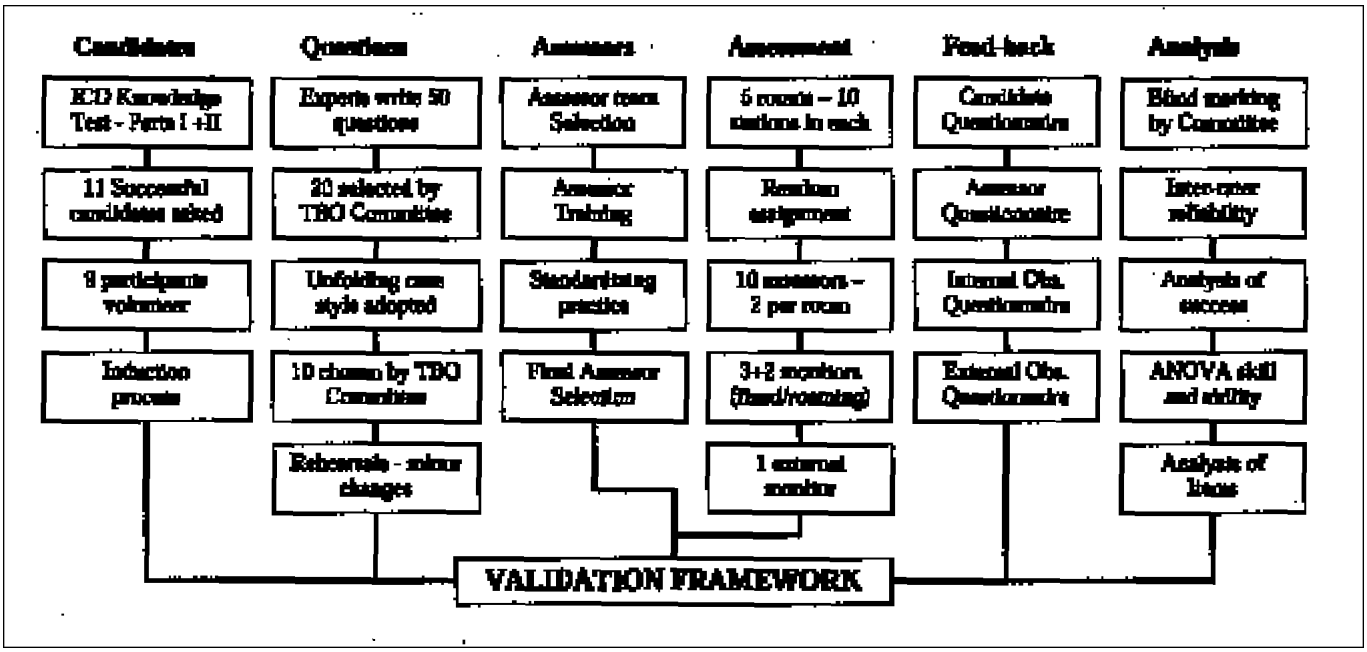

Fig. 1 - *Chart of the validation framework for the objective structured clinical examination pilot.*

procedural conditions and researching the impact a substantial contribution can be made to validating an examination, a complex task (8), and therefore to establishing its ability to measure the traits its sets out to measure.

MATERIALS AND METHODS

The test process was initiated and followed through by the Assessment Committee of the Turkish Board of Ophthalmology (TBO), which planned and oversaw the development of the pilot study following the validation process outlined in Figure 1. In this section the procedures used for the construction and selection of items and administration of the test are outlined, followed by a review of the arrangements for assessor selection and training. The scheme adopted for examinee selection is then presented, with an account of the methods used to ensure and monitor reliability and validity of the test.

OSCE test construction and administration

A bank of 50 questions, based on the Turkish Ophthalmology Society Education Committee's global minimum essential requirements curriculum, was generated by a selected group of experts, who were different from the assessors involved in the actual assessment. These were narrowed down to 20 questions by the TBO Assessment Committee. The questions were then reformatted by one of the committee members (P.A.) into unfolding cases in the OSCE style (9) and subjected to further analysis and selection by being read aloud to the Assessment Committee who offered comments, made adjustments accordingly, and came up with a final set of 10 questions. Finally, two on stage rehearsals were performed during an assessor training and standardization session, discussed below, which again submitted the items to scrutiny. Based upon the assessors' feedback, and subject to the agreement of the Assessment Committee, a few minor changes were made to the checklist items.

The total time allocated for the examination for each candidate was precisely 90 minutes, namely 10 minutes for each question, bar two questions for which 5 minutes each were given. Candidates were assigned to one of five rooms for the administration of the examination. Each room contained 10 stations, one for each question, which candidates visited in turn and where they answered the associated question and subquestions, in the unfolding case style. Two assessors were assigned to each room, following the candidate around the stations, with the whole process monitored by five observers from the TBO Assessment Committee, one in each of three of the rooms and two roaming. Assessors were asked to go over all questions in the same time period and the

process was regulated by the ringing of a bell, heard by all assessors at the same time, which announced the end of one question and the beginning of the next. Each candidate had an instruction sheet giving instructions of each step for each station. Assessors had a checklist which included instructions for candidates and a checklist for potential answers, as well as a detailed scenario for the role playing question in which one assessor had to participate as an interlocutor to the candidate. During the examination assessors filled in the documentation provided and submitted the candidate's performance results, without collusion, directly to the Assessment Committee. No feedback was given during the assessment by the assessors, and no guidance other than that designated in the examination documents was allowed. Neither candidates nor assessors were allowed to take away any of the documentation.

The attribution of marks to performance was carried out independently by Assessment Committee members on completion of the examination. Final marking of all examples of test performance was carried out individually and blind by each of the five members of the TBO Assessment Committee: each candidate's output was marked independently five times and the final mark given was an average of the five independent assessments. Each of the questions was accorded the same mark–10 marks out of 100–but the distribution of the marks under each question depended on the predetermined weighting given to subitems. Fundus drawings were also marked blind by all committee members, using a checklist of criteria set prior to the assessment.

In line with the OSCE model (1-7), the examination was designed around two main aims: first, to assess performance skills, i.e., a candidate's capacity to perform an ophthalmologic examination or procedure on patients; and, secondly, to assess the ability to manage patients, i.e. the ability to problem-solve related to prevention, diagnosis, treatment, and follow-up of patients. Knowledge-based questions were avoided as much as possible. Skill and abilities were not separated during the test but were integrated in the checklists used by the assessors. Thus, candidates were assessed on a holistic performance reflecting a real-life clinical situation, with the breakdown of performance into skills or abilities achieved through the checklist rubric. The following scenarios were used as the basis for the test of skills and abilities: a refraction examination performed on a phantom eye with streak retinoscope; fundus examination and schematiza-

TABLE I - PERCENTAGE SCORES FOR ABILITY AND SKILL TAKEN SEPARATELY FOR EACH CANDIDATE

Candidates	Fundus	Uveitis	Glaucoma	Operation	Refraction	Strabismus	Trauma	Exophthalmia	Bullous	Tumor	Mean
S3	86	100	42	100	67	100	100	86	75	100	86
S4	86	100	75	100	83	75	50	100	75	83	83
S1	93	100	75	67	50	94	75	82	75	100	81
S8	43	90	75	100	33	100	94	100	75	100	81
S2	93	70	83	100	67	88	88	82	50	83	80
S5	71	80	33	100	100	100	75	82	50	83	77
S6	71	80	67	100	33	100	75	100	50	67	74
S7	68	80	17	100	67	85	68	100	50	83	72
S9	68	80	50	33	0	100	75	100	25	67	60
Skill mean	**75**	**87**	**57**	**89**	**56**	**94**	**78**	**92**	**58**	**85**	**77**
A7	100	55	94	100	0	67	75	100	83	100	77
A8	75	70	83	100	50	54	63	100	67	100	76
A2	75	85	86	36	50	67	88	100	67	100	75
A3	100	80	93	64	0	67	79	100	67	100	75
A5	50	70	57	64	100	58	92	100	50	100	74
A6	100	85	86	71	0	42	100	100	50	100	73
A4	100	50	93	71	0	54	75	100	50	100	69
A9	75	55	86	14	0	63	75	100	83	100	65
A1	50	80	82	61	0	67	42	50	50	100	58
Ability mean	**81**	**70**	**84**	**65**	**22**	**60**	**77**	**94**	**63**	**100**	**72**

tion on a phantom eye, with indirect ophthalmoscope and FFA interpretation on a related FFA slide; history taking in uveitis using a guided role play; history taking in glaucoma based on a patient whose disc photograph and visual field results were shown; surgery complication assessment and prevention; eye movements examination on a nine-direction eye movement photograph of a patient using a slide; eye examination in a case with anterior segment trauma using a slide; exophthalmos examination on one of the assessors with exophthalmometer; recognition of a lid tumor using a slide; prescription writing for a patient with a common cornea problem.

Assessor selection and training

The TBO Assessment Committee invited potentially interested assessors from university ophthalmology residency education programs from different parts of the country. A total of 12 assessors eventually applied and were selected to take part in the pilot assessment project. As none of them had ever sat an OSCE, nor taken part as assessors in this format of assessment, a 2-hour theoretical and practical induction was given just prior to the examination, covering mission and goals; OSCE type assessment; assessor roles and responsibilities; candidate question sheets; and the checklist sheets, in written and verbal form. Assessors were then introduced to the examination rooms and the assessment tools, i.e., streak retinoscope, indirect ophthalmoscope, and exophthalmometer, and went through a practice rehearsal for the examination itself. During this process they were asked to express their views openly about the checklist items and to consider the current examination as a trial of these. Assessors were also asked to fill out a form after the examination, of a type similar to that given to the candidates, as discussed below.

Candidate selection and induction

From 2003 onwards the TBO actively encouraged ophthalmology specialists in Turkey to take the knowledge assessment examination offered by the International Council of Ophthalmology (ICO) Assessment Committee (http://www.icoph.org/assess/index.html). The TBO Assessment Committee invited ICO candidates who had achieved a pass in this two part assessment (basic and clinical) to take part on a voluntary basis in the first TBO-OSCE pilot. Nine of the 11 ICO pass candidates who applied took part in the pilot OSCE assessment, four female and five male, aged between 27 and 40 years (mean: 34). They had graduated as specialists from different ophthalmology residency programs from various parts of the country between 2 and 15 years earlier (mean: 6.2). Under Turkish law they were all qualified ophthalmic surgeon-consultants, authorized to manage an ophthalmology clinic and carry out ophthalmic surgical procedures. Thus,

TABLE II - COMPARISON OF CANDIDATE (N=9) AND ASSESSOR (N=12) QUESTIONNAIRE RESPONSES

Overall difficulty level	**Easy**	**Moderate**	**Difficult**
Candidates (out of 90)	39 (43.3)	40 (44.5)	11 (12.2)
Assessors (out of 120)	51 (42.5)	54 (45)	15 (12.5)
Overall success in determining knowledge level	**Weak**	**Moderate**	**Strong**
Candidates (out of 90)	49 (54.5)	36 (40)	5 (5.5)
Assessors (out of 120)	56 (46.7)	52 (43.3)	12 (10)
Overall success in assessment of skills	**Weak**	**Moderate**	**Strong**
Candidates (out of 90)	17 (18.8)	48 (53.5)	25 (27.7)
Assessors (out of 120)	18 (15)	45 (37.5)	57 (7.54)
Overall success in assessment of ability	**Weak**	**Moderate**	**Strong**
Candidates (out of 90)	17 (18.8)	36 (40)	37 (41.2)
Assessors (out of 120)	19 (15.8)	32 (26.6)	69 (57.6)
Overall duration	**Short**	**Good**	**Long**
Candidates (out of 90)	1 (1.1)	61 (67.8)	28 (31.1)
Assessors (out of 120)	10 (8.3)	48 (40)	62 (51.7)
Overall success of the assessment	**Poor**	**Medium**	**Good**
Candidates	0	2	7
Assessors	0	1	11

Values are n (%)

they had all been trained as ophthalmic surgeons with their surgical skills monitored by means of a log book during their residency, a common practice in the country and elsewhere (http://www.ebo-on-line.org).

As none had ever taken an OSCE before, a 2-hour induction program on the aims, format, and duration of the question groups was given prior to the examination. In a similar way to the assessors, all examinees were asked to fill in a feedback form in which they were asked to rate the overall difficulty level on a scale of easy-medium-difficult; to rate effectiveness of knowledge assessment, effectiveness of skill assessment, and effectiveness of ability assessment on a scale of weak-medium-strong; to rate the duration of the examination on a scale of short-good-long; and to provide an overall success rating of the assessment on a scale of poor-medium-good.

Validation process

In order to ensure reliability of the examination certain measures were taken. First, a process of insider observation was initiated. Three members of the Assessment Committee acted as insider observers and were attached individually to a single room during the whole examination process, whereas two others were chosen to observe the assessments randomly, moving from room to room at will. Secondly, an independent outsider observer, experienced in undergraduate OSCE, was invited to observe the administration of the test. After the observation, both the independent observer and the insider observers were asked to fill out separate questionnaires, prepared prior to the assessment, which included questions concerning mainly the flow of the procedure, the setting, and reliability factors.

TABLE III - COMPARISON OF FEEDBACK BETWEEN INTERNAL (N=5) AND EXTERNAL (N=1) OBSERVERS

Difficulty level of the questions	**Easy**	**Medium**	**Difficult**
Overall internal observers	3	2	0
External observer	No comments	—	—
Duration of the questions	**Short**	**Medium**	**Long**
Overall internal observers	0	1	4
External observer	No comments	—	—
Overall assessors/candidate ratio	**Few**	**Good**	**Much**
Overall internal observers	1	3	1
External observer	—	1	—
Amount of tools	**Few**	**Good**	**Much**
Overall internal observers	1	4	0
External observer	—	1	—
Environment of the exam setting	**Not adequate**	**Adequate**	
Overall internal observers	1	4	
External observer	—	1	—
Induction of the assessors	**Too little**	**Good**	**Too Much**
Overall internal observers	1	4	0
External observer	1	—	—
Induction of the candidates	**Too**	**Little Good**	**Too Much**
Overall internal observers	2	3	—
External observer	1	—	—
Flow (quick adjustment to next question)	**No problems**	**Minor problems**	**Major problems**
Overall internal observers	—	5	—
External observer	—	1	—
Station setting style	**Good in same room**	**Try Different rooms**	
Overall internal observers	1	4	—
External observer	—	1	—
Assessors cooperation	**Poor**	**Medium**	**Good**
Overall internal observers	0	2	3
External observer	—	—	1
Candidates cooperation	**Poor**	**Medium**	**Good**
Overall internal observers	—	3	2
External observer	—	—	1

As mentioned above, measures were taken to enhance scorer reliability prior to the examination by induction of assessors, and a rehearsal and standardization process, and after the examination by blind marking and the separation of those assessing student product and calculating final scores. In addition a small scale verification of inter-rater reliability was carried out in one of the rooms through asking the two assessors and the monitor to fill in, and not to compare, separate checklists for each candidate they were assessing. These checklists were then compared following the examination with a view to ascertaining the level of variability and thus the effectiveness of the standardization process.

Validity of the examination, of which reliability is a factor (8), was monitored through the questionnaires given to the participants, the assessors, and the observers, with all 9 candidates, 12 assessors, 5 internal observers, and 1 external observer providing completed questionnaires. The fact that candidates represented a homogenous group also contributed to the validation process. Their previous background meant that they were already at a level susceptible to passing the examination. Hence, assessment criteria were being applied to a relatively homogenous population with potentially similar skill levels, thus testing the ability of the examining process to accurately differentiate between relatively small expected variations in the level of candidate performance.

RESULTS

Candidate assessment results

The candidate success rate on the examination was 89% with the pass mark set at 70%. Based on feedback all main question areas were considered to be indispensable for success in the assessment and none were eliminated. However, four subitems from the marking checklists were deleted as none of the candidates had displayed them. Table I separates skill performance elements in the 10 questions from those focused on ability and converts the scores into percentages. The figures indicate that, in the main, candidates revealed themselves more successful in the skill elements in questions (mean 77%) than in clinical ability (mean 72%). Three areas–operation, refraction, strabismus–indicate substantial differences of over 20% in performance between skill and ability, whereas only glaucoma shows an inverse trend. However, in general, skill and ability performance differences are correlated ($r=0.41$) and non-significant on a post hoc, two-way analysis of variance (ANOVA). There are, on the other hand, significant differences, based on the ANOVA, in performance between the different questions ($p<0.05$).

Post-examination item analysis is a process which uses the spread of answers between the more successful and less successful examination takers to decide whether question construction or difficulty level of items in an examination is at an acceptable level. Table I ranks the candidates based on their scores from highest to lowest and provides a picture of which items were easily answered by all, and which were more problematic. This is generally expressed statistically by a facility index, namely, in this case the number who passed as a proportion of total takers. The ranking can also indicate if the construct underlying the question was problematic or not by showing whether the item was answered correctly by the highest scoring candidates as compared to the lowest scoring candidates, generally expressed statistically by a point biserial correlation coefficient. In the examination here it appears that certain items could be considered easy, for example the ability assessment of exophthalmia or tumor, or certain items inconsistent, for example the skill assessment of exophthalmos in which the lowest overall scoring candidates got the highest marks.

Candidate impressions (questionnaire results)

All candidates were positive about the explanations given in the induction prior to the assessment, as well as the assessment performance requirements, including the timing and question styles.

All but one mentioned that they enjoyed the experience, the odd one out being the person who failed to meet the pass standard. The analysis of the feedback questionnaires in Table II is based on responses to each question by all candidates ($n=9$) compared to all assessors ($n=12$). It shows that none of the candidates had felt threatened by the assessment format.

Two mentioned that they would have much preferred questions concerning treatment modalities rather than goal of treatment or patient communication.

One suggested that it would have been easier to have multiple choice questions (MCQ) rather than a test of skill and ability. The overall difficulty level was found easy to medium by candidates; the effectiveness of knowledge assessment was found to be weak; the effectiveness of

skill assessment was found to be strong to moderate; the effectiveness of the ability assessment was found to be moderate to strong; the duration of the examination was found to be good; and the general assessment of the procedure was found to be good. All emphasized that OSCE examination types are useful and necessary.

Assessor impressions (questionnaire results)

The initial reaction of the assessors could be described as skeptical, which was resolved through the training session. Their second reaction was to be critical of question construction, which was resolved by making minor changes in the checklist forms. Their third reaction turned out to be positive, shown through their willingness to join in the process, and their final reaction was to demonstrate concern to carry out their task without mistake and within the rules. Faculty members claimed that they enjoyed the assessment format as it forced them to think about what and how they had been teaching the skills and abilities covered in the examination. It also forced them to reflect on how they would have performed had they been taking the examination as a candidate. They also stated that they would be trying to use a similar format for undergraduate students in their institutions, an indication of some early backwash effect. Overall, they stated they had been happy to take part in the procedure and would willingly do so again. The feedback (Tab. II) indicates that the overall difficulty level was found to be easy to moderate; the effectiveness of knowledge assessment was found to be weak; the effectiveness of skill assessment was found to be strong; the effectiveness of the ability assessment was found to be very strong; the duration of the examination was found to be long; and the general assessment of the procedure was found to be good. Everyone believed that OSCE style examination was both useful and necessary.

Inside and outside observer impressions (questionnaire results)

The experienced outsider observer visited the rooms, checked the question settings, question sheets, and checklists, as well as the feedback forms for both candidates and assessors. According to this external observer, the assessment setting, including station number and candidate/assessor ratio, were optimal. No practical problem which could have interfered with the flow of the assessment occurred during the examination process (Tab. III).

Inter-rater reliability

High correlations ($r>0.99$) were found between the two assessors and one monitor who filled in separate checklist forms for the same candidate, which indicates that the three persons concerned had internalized the training, and that, in general terms, reasonable reliability estimates can be achieved for this type of examination. In addition no significant difference was found in the blind scoring of papers by the five members of the Assessment Committee.

DISCUSSION

Conventional MCQ assessment has proved to be a quick and reliable means of accurately assessing knowledge (10). On the other hand, in clinical specialties which require surgery, these conventional assessment types are not able to provide a valid assessment of clinical examination skills, nor can they adequately assess the appropriate approach to adopt to patients in a variety of clinical situations (1-3, 11). It is possible to assess surgical skills by direct supervision through a residency program, monitoring progress using log books (http://www.ebo-online.org). Likewise, it is possible to assess clinical skills by direct observation and oral assessments. However, the disadvantage of these assessment types is both their lack of apparent objectivity due to the problem of comparability between different contexts and also the amount of time they consume. An OSCE examination procedure can potentially eliminate these disadvantages by the use of performance-based assessment formats which measure clinical skills and patient handling competence through observation, in a relatively short period of time (1-4). Thus, an OSCE procedure provides an independent measure of surgical skills and patient handling abilities, tied to an externally defined standard.

The intent of the examination appears to have been recognized by candidates and assessors through their perceptions of the knowledge, skill, and ability requirements of the questions. This indicates a good fit between the aims of the examination–to test skill and ability–and the perception of the candidates of the examination question types. The acceptance of the OSCE format by the candidates, shown through the results from the questionnaire feedback, reveals that an OSCE type of assessment is capable of being well accepted and trusted by candidates. Trust is an important indicator of face validity for a

test. The fact that some of the candidates wished to have questions on the treatment of diseases, rather than the skills concerned with detection and dealing with patients, is perhaps indicative of the need to review take-away forms of medical education. Similarly, the fact that some candidates would have preferred MCQs shows the impact of traditional forms of testing on education. Test results indicate, however, that candidates have significant gaps in their skills and abilities, subject to the assumption that all questions have been well designed, i.e., have construct validity, and thus differences in answers can be attributed to variations in skill and ability. In addition, the analysis of answers to items shows that there is no rank correlation ($r=-0.16$) between the candidates' performance on the skills part and the ability part of the assessment. This would tend to give weight, based on this small trial, to the need for tests that focus on both areas–skills and abilities–in that possession of competence in one area does not appear to predict performance in the other.

The analysis of questionnaire results in the study indicated that the amount of time allocated to each question and to the examination as a whole was considered too long by assessors and observers, echoed to some extent by the examinees. Finding a balance for this type of assessment is clearly a challenge where the need to have adequate coverage of the curriculum has to be matched by efficient methods, particularly in large scale testing in which cost might be an issue. Both the assessors and the candidates considered that over 50% of questions were moderate to difficult, showing an overall reasonable spread of difficulty, a desirable characteristic. The results of the analysis of examinee responses to items, on the other hand, indicated that some items were too easy or had given inconsistent results. As the questions were criterion referenced, any doubts raised as to the level of difficulty or the construct under lying a question would need to be resolved through a reconsideration of the tasks by an expert panel in order to feel confident that the task descriptions were valid and reflected the required knowledge for the level of the examination.

A weakness in the study was the low number of candidates in the trial, which meant that the candidate-assessor ratio of one to three turned out to be unnecessarily high. The holding of rehearsals for assessors prior to the examination was an important aspect of ensuring the success of the assessment as it standardized procedures and gave confidence to assessors who had never experienced this format of assessment before. The initial doubts expressed by the assessors were successfully eliminated through allowing them to express their views and, accordingly, make minor alterations to the items in the checklist. This is an issue which would require further consideration in larger trials, or where the examination was run in multiple centers, as such changes to checklists, for example, might introduce unreliability through modifying examination conditions or standards across centers.

A further area of weakness of the current pilot design was the lack of inter-rater reliability measures across the five examination rooms. The high correlation ($r>0.99$) established among the three assessors who filled in separate checklists for the same candidate shows that checklist-forms can be used reliability to assess performance. This indicated that the assessor training had been effective and, by extension, that some confidence could be expressed as to the effectiveness of the other assessors in the other rooms. It would, however, have been desirable to extend the rater-reliability study to all assessors so that between-room comparisons could have been made. In examinations of this type involving large numbers of candidates, statistical techniques such as three parameter Rasch analysis (14) can be used to estimate rater reliability and, eventually, eliminate those raters who do not perform consistently. However, on the basis of experience here, it might also be desirable to train selected raters to mark all candidates at one task rather than have all raters assess all candidates at all tasks, thus ensuring a higher likelihood of consistency.

Real cost considerations (12) have not been to the fore in the trial described here. However, in large scale testing contexts such costs would inevitably have an impact on the examination and would require consideration. Costs involve quality issues and vary positively as a function of the dependability of the examination, the assessors involved, and the mode of operation. For example, the use of standardized patients in order to increase reliability (13) would inevitably introduce extra cost considerations, and these would have to be offset against the gains in efficiency to be achieved. In the long term an optimum balance between costs in terms of efficiency versus effectiveness needs to be found.

The study reported above has developed a validation framework for the OSCE examination and goes some way to ensuring that different types of validity are addressed (8), a concern for all serious examining bodies. The study has attempted to ensure that the examination exhibits face validity through sharing of perspectives between takers and

administrators of the examination. In addition, it has shown the importance of the training of perceptions, particularly for the examiners. The study has also emphasized the crucial role of a team of specialist question writers to agree on what constitutes acceptable criterion related measures of skills and abilities. The right sample of behaviors, abilities, and skills to be examined allows a confident prediction of the candidate's ability. Furthermore, the post-examination analysis of the questions ensures that the sample is at the right level and well constructed and, taken together with expert opinion, provides a strong indication of a test's validity. A possible next step in establishing validity of the OSCE could be the observation of candidates in their daily practice to see if the examination results reflect their performance in real-life conditions.

This might provide a useful indication of the impact of factors such as examination nerves on candidates, and also furnish extra data to justify the inclusion of certain skills and abilities in the assessment criteria.

In conclusion, the pilot study shows that OSCE style assessment can be applied successfully to postgraduate, surgically oriented specialties such as ophthalmology and provides an effective means of making judgments as to skills and abilities of candidates. Indications of some examination backwash on assessors' practice are a sign that this form of assessment will have a positive impact on the implementation of the curriculum of independent education clinics nationwide. The study leads us to believe that this form of assessment will gain in national recognition, but more pilot studies on a larger scale, within an improved validation framework, are needed in order to promote this.

ACKNOWLEDGEMENTS

The authors thank the assessors who participated in the study and John O'Dwyer, PhD, who reviewed the study from an educational assessment perspective.

The authors have no proprietary interest.

Reprint requests to:
Pinar Aydin, MD, PhD
Tunali Hilmi Caddesi No: 110/13
Kavaklidere, Ankara, 06700, Turkey
aydinpinartr@yahoo.com

REFERENCES

1. Cohen R, Reznick RK, Taylor BR, Provan J, Rothman A. Reliability and validity of the objective structured clinical examination in assessing surgical residents. Am J Surg 1990; 160: 302-5.
2. Sloan DA, Donnelly MB, Johnson SB, Schwartz RW, Strodel WE. Use of an objective structured clinical examination (OSCE) to measure improvement in clinical competence during the surgical internship. Surgery 1993; 114: 343-50; discussion 350-1.
3. Sloan DA, Donnelly MB, Schwartz RW, Strodel WE. The objective structured clinical examination. The new gold standard for evaluating postgraduate clinical performance. Ann Surg 1995; 222: 735-42.
4. Sloan DA, Donnelly MB, Schwartz RW, Felts JL, Blue AV, Strodel WE. The use of objective structured clinical examination (OSCE) for evaluation and instruction in graduate medical education. J Surg Res 1996; 63: 225-30.
5. 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 [erratum in 2003; 37: 574]. Med Educ 2003; 37: 132-9.
6. Kramer AW, Jansen KJ, Dusman H, Tan LH, van der Vleuten CP, Grol RP. Acquisition of clinical skills in postgraduate training for general practice. Br J Gen Pract 2003; 53: 677-82.
7. Kramer AW, Jansen JJ, Zuithoff P, Dusman H, Tan LH, Grol RP, van der Vleuten CP. Predictive validity of a written knowledge test of skills for an OSCE in postgraduate training for general practice [erratum in 2003; 37: 83]. Med Educ 2002; 36: 812-9.
8. Weir CJ. Language Testing and Validation: An Evidence-based Approach. London: Palgrave MacMillan; 2004; 11.
9. Karani R, Callahan EH, Thomas DC. An unfolding case with a linked OSCE: a curriculum in inpatient geriatric medicine. Acad Med 2002; 77: 938. Review.
10. Keely E, Myers K, Dojeiji S. Can written communication skills be tested in an objective structured clinical examination format? Acad Med 2002; 77: 82-6.
11. Rymer AT. The new MRCOG Objective Structured Clinical Examination–the assessor's evaluation. J Obstet Gynaecol 2001; 21: 103-6.
12. Reznick RK, Smee S, Baumber JS, et al. Guidelines for estimating the real cost of an objective structured clinical examination. Acad Med 1993; 68: 513-7.
13. Regehr G, Freeman R, Robb A, Missiha N, Heisey R. OSCE performance evaluations made by standardized patients: comparing checklist and global rating scores. Acad Med 1999; 74 (10 Suppl): S135-7.
14. Bond TG, Fox CM. Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahwah, NJ: Lawrence Erlbaum Associates; 2001; 106-18.