A Modified OSCE Assessing the Assimilation and Application of Ethical Principles Relevant to Obstetric and Gynaecological Practice

Dr H van Woerden, Dr F Agbo*, Mr NN Amso†, Mr I Stokes‡

*Senior House Officer, Obstetrics and Gynaecology
University Hospital Wales, Cardiff, CF14 4XN.

†Senior Lecturer/Honorary Consultant,
University of Wales College of Medicine
and the University Hospital of Wales, Cardiff, CF14 4XN.

‡Consultant Obstetrician and Gynaecologist,
Nevill Hall Hospital, Abergaveny, NP7 7EG.

Abstract: Objective. To develop and evaluate a modified OSCE assessing the assimilation and application of a range of ethical principles relevant to Obstetric and Gynecological practice.

Setting. Candidates for an SpR training rotation

Methods. Twenty six candidates working in Obstetrics and Gynecology were presented with four questions covering a range of relevant ethical scenarios. Their responses were assessed using a marking schedule. The marking schedule was evaluated against a checklist developed for assessing postgraduate medical examinations. Inter-rater reliability was assessed by calculating Kappa values for each question. The items in the marking schedule were also assessed to determine the level of agreement between the two examiners. To assess the contribution of each question to the total score, the question to total score correlations were calculated. The discriminatory capacity of each question was also assessed.

Results. The development of the examination met almost all of the criteria in the checklist for developing a postgraduate examination. Inter-rater reliability was reasonable (4 weighted Kappas ranged from 0.53 - 0.75). There was a high level of agreement between examiners as to whether a candidate had answered an item on the marking schedule correctly. The degree of discrimination of items in the marking schedule was consistent with clinical opinion on the importance of questions.

Conclusion. This modified OSCE examination demonstrates the feasibility of testing ethical principles relevant to practice in Obstetrics and Gynecology in candidates for postgraduate posts. It meets most of the criteria laid down in a checklist developed to assess postgraduate medical examinations.

Keywords: OSCE; examination; ethics; Obstetrics and Gynecology.

A modified form of the Objective Structured Clinical Examination (OSCE) has become a standard part of the Part II of the Obstetrics and Gynecology membership examination in Wales. It is a useful method of examining knowledge and practice across a set of clinical, administrative and ethical areas of competence. This modified OSCE involved the use of an examiner instead of an actor to simulate the patient.

The modified OSCE developed for this study was designed to assess assimilation and application of the principles outlined in Royal College of Obstetricians and Gynecologists’ (RCOG) guidelines. These cover the following areas: general attitude to women, consent to treatment and examination by medical students, clinical training, use of tissue, research, innovative procedures and professional disagreement.

Methods

Four questions were developed which translate the principles of the RCOG Guidelines on ethical
practice into a modified OSCE. Two questions applied the principles in a clinical context, one question in the area of clinical governance and one question in an area of non-clinical practice (Appendix 1). Questions were developed by one of the authors (NNA). The modified OSCE was used on two sets of 10 and 16 candidates being interviewed for entry to a Specialist Registrar training scheme in obstetrics and gynecology in South Wales in 2000.

All candidates were Senior House Officers in Obstetrics and Gynecology looking to obtain Registrar posts and therefore had similar levels of experience. Candidates were given 15 minutes to read the RCOG Ethical Guidelines\(^3\), they were then interviewed by two doctors, (NNA and IS) for 15 minutes. All 26 candidates were seen by the same two examiners. Prior to the start of the examination, discussion took place between the two examiners as to the interpretation of specific items on the marking schedule. However, each examiner was blind to the marking of the other examiner during the interview process.

The marking schedule was assessed retrospectively using a set of criteria proposed for the assessment of postgraduate medical examinations.\(^4\)

Each item on the marking schedule was scored 0 if no attempt was made to address the relevant concept, 1 if a partial answer was provided and 2 if the item was fully addressed. Marks were entered in a Microsoft Excel\(^\text{TM}\) spreadsheet.

The mean and standard deviation of scores given to candidates by each examiner was calculated and plotted in two histograms using the Statistical Package for the Social Sciences (SPSS\(^\text{®}\)).

The level of agreement or disagreement between the two examiners was measured by calculating the average difference between examiners in their assessment of whether candidates had addressed the issue sought in that item. If both examiners scored a candidate as 2, 1 or 0, the “difference between examiners” was 0 (i.e. no difference). If one examiner score a candidate as 2 and the other as 1, or if the first examiner scores a candidate and 1 and the second as 2, the “difference between examiners” was 1 (i.e. moderate). If one examiner scores a candidate as 2 and the other as 0, the “difference between examiners” was 2 (i.e. large). The average of the absolute difference between examiners for each candidate was then calculated.

Inter-rater reliability was also assessed by calculated weighted and unweighted Kappa statistics for each of the four questions using Pepi 30 software.

To assess the contribution of each of the four questions to the total score, correlation were calculated. Correlation values were calculated in Excel\(^\text{TM}\) using the average score of the two examiners for each item in the marking schedule and correlating this with the total score of each candidate.

The ability of each question to discriminate among the candidates was assessed by examining the proportion of candidates who either scored full marks on each score sheet item or did not score on that item at all.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figures.png}
\caption{A normal distribution is indicated by the superimposed continuous line.}
\end{figure}
Results

Data from the marking schedule of one candidate by one of the examiners was missing. This resulted in some of the analyses including only 25 candidates. The one marking schedule which was available on this candidate did not indicate that these marks were in any way atypical.

The mean, standard deviation and a histogram of the scores given to the candidates by each examiner are shown in Figure 1. The level of agreement between examiners for each part of the four questions is shown in Table 1.

Weighted and unweighted Kappa values and the correlation of each item with the total score are shown in Table 2. Table 3 presents the frequency of responses for the three rating categories for each item in the marking schedule. This provides a means of assessing the ability of each item to discriminate among the examinees. McNamar test for bias gave a p=1.00 for all four items suggests that there was no consistent bias by one or other of the examiners.

The results of assessing the marking schedule against the set of criteria proposed for the assessment of postgraduate medical examinations is outlined in Table 4.

Table 1. Level of agreement between examiners

| Question | Examiners agree/ partly agree/ disagree | Average difference between examiners scores | Question | Examiners agree/ partly agree/ disagree | Average difference between examiners scores |
|----------|-----------------------------------------|---------------------------------------------|----------|-----------------------------------------|---------------------------------------------|
| Q1.1     | 19/ 4/ 1                                | 0.28                                        | Q3.1     | 22/ 2/ 2                                | 0.16                                        |
| Q1.2     | 21/ 3/ 1                                | 0.20                                        | Q3.2     | 17/ 7/ 0                                | 0.32                                        |
| Q1.3     | 21/ 2/ 1                                | 0.20                                        | Q3.3     | 23/ 2/ 1                                | 0.08                                        |
| Q1.4     | 20/ 3/ 1                                | 0.24                                        | Q3.4     | 22/ 3/ 0                                | 0.12                                        |
| Q1.5     | 19/ 4/ 1                                | 0.32                                        | Q3.5     | 23/ 2/ 1                                | 0.08                                        |
| Q2.1     | 24/ 1/ 0                                | 0.04                                        | Q3.6     | 22/ 3/ 0                                | 0.12                                        |
| Q2.2     | 21/ 3/ 1                                | 0.20                                        | Q3.7     | 21/ 0/ 4                                | 0.32                                        |
| Q2.3     | 23/ 0/ 2                                | 0.16                                        | Q4.1     | 24/ 1/ 0                                | 0.04                                        |
| Q2.4     | 24/ 0/ 1                                | 0.08                                        | Q4.2     | 19/ 5/ 1                                | 0.28                                        |
| Q2.5     | 18/ 1/ 4                                | 0.48                                        | Q4.3     | 22/ 3/ 0                                | 0.12                                        |
| Q2.6     | 14/ 8/ 2                                | 0.52                                        | Q4.4     | 21/ 4/ 0                                | 0.16                                        |
| Q2.7     | 19/ 5/ 0                                | 0.24                                        | Q4.5     | 11/ 10/ 3                               | 0.68                                        |
| Q2.8     | 24/ 1/ 0                                | 0.04                                        | Q4.6     | 20/ 4/ 1                                | 0.24                                        |
| Q2.9     | 19/ 5/ 0                                | 0.24                                        | Q4.7     | 12/ 9/ 3                                | 0.64                                        |
| Q2.10    | 22/ 3/ 0                                | 0.12                                        | Q4.8     | 13/ 8/ 4                                | 0.64                                        |

Key
* Almost complete agreement between examiners as to whether candidates have address this issue.
† Lowest agreement between examiners as to whether candidates have address this issue.

Discussion

An OSCE approach has previously been used in assessing other aspects of obstetrics and gynaecology, and in testing knowledge of ethical aspects of clinical practice, however no previously published OSCE could be identified which tested the assimilation and application of ethical principles in the practice of obstetrics and gynecology. We were also unable to identify the previous use of an OSCE to select candidates for a postgraduate training scheme.

Figure 1 demonstrates the similarity in the characteristics of the two examiners ratings apart from the fact that Examiner 2’s marks have a wider standard deviation and a different modal value. Examiner 2 also appears to dichotomize candidates’ marks to produce a more binomial distribution. This may simply be a random effect due to a small sample size. However, it is possible that this reflects a different marking style. Examiner 1 gave most candidates an average mark and a small number of candidates high or low marks. Examiner 2 may subconsciously have classified most candidates as falling into two groups: stronger candidates who would pass and weaker candidates who would fail. This hypothesis would require confirmation in a larger sample of examiners. The use of a wider sample of examiners would also
address other issues including the effect of the gender of an examiner and any bias that may have been introduced by using one of the authors as an examiner.

The average absolute difference between the raters for each item was at most 0.68 and for most items substantially less. This suggests that there is good agreement between the examiners in their marking of candidates. However, it should be noted that two sets of independently, randomly assigned values of 0, 1, and 2 have a 'chance level of disagreement' of about 0.9. Consequently, the result we obtained for the level of agreement between the examiners should be interpreted with caution. Inter-rater reliability was also measured using the Kappa statistic. Based on the weighted Kappa values, the inter-rater reliability were good for questions 1, 2 and 3 and moderate for question 4 suggesting the procedures used in the examination allow for reasonable inter-rater reliability.

Six items displayed relatively low discrimination while three had a high level of discrimination. Items 3.6, 4.3 and 4.4 appear difficult; however, there were no questions which all of the examinees missed.

The scale used in the marking schedule is categorical and may not reflect a true interval scale. The validity of adding and averaging of such scores is potentially suspect and other methods of quantification of responses could be considered if this instrument was used in other situations.

The correlation between question scores and the total score (Table 2) appears to be in part related to the number of concepts included under each question with the questions encompassing more areas exhibiting a higher correlation with the total score. An alternative approach to scoring the exam, which addresses this issue, would involve weighting each

### Table 2. Kappa values for the four questions

| Question | Kappa | Weighted Kappa | 95% CI for Weighted Kappa | Correlation with total score |
|----------|-------|----------------|---------------------------|-----------------------------|
| 1        | 0.39  | 0.75           | 0.63 - 0.87               | 0.49                        |
| 2        | 0.29  | 0.62           | 0.43 - 0.82               | 0.69                        |
| 3        | 0.32  | 0.65           | 0.49 - 0.81               | 0.44                        |
| 4        | 0.17  | 0.53           | 0.38 - 0.68               | 0.53                        |

### Table 3. Percentage of ratings in each category of the items in the marking schedule

| Marking Schedule Item | Not addressed | Partly addressed | Fully addressed | Marking Schedule Item | Not addressed | Partly addressed | Fully addressed |
|-----------------------|---------------|------------------|-----------------|-----------------------|---------------|------------------|-----------------|
| 1.1 †                 | 28%           | 24%              | 48%             | 3.1 *                 | 0%            | 8%               | 92%             |
| 1.2                   | 20%           | 16%              | 64%             | 3.2 †                 | 56%           | 16%              | 28%             |
| 1.3                   | 20%           | 12%              | 68%             | 3.3 *                 | 4%            | 4%               | 92%             |
| 1.4 †                 | 36%           | 28%              | 36%             | 3.4                   | 64%           | 32%              | 4%              |
| 1.5 †                 | 28%           | 28%              | 44%             | 3.5 †                 | 56%           | 4%               | 40%             |
| 2.1 *                 | 4%            | 4%               | 92%             | 3.6                   | 88%           | 8%               | 4%              |
| 2.2 †                 | 60%           | 16%              | 24%             | 3.7 †                 | 48%           | 4%               | 48%             |
| 2.3                   | 12%           | 0%               | 88%             | 4.1                   | 12%           | 12%              | 76%             |
| 2.4                   | 16%           | 0%               | 84%             | 4.2 †                 | 52%           | 20%              | 28%             |
| 2.5 †                 | 44%           | 16%              | 40%             | 4.3                   | 88%           | 8%               | 4%              |
| 2.6 †                 | 52%           | 20%              | 28%             | 4.4                   | 84%           | 12%              | 4%              |
| 2.7                   | 80%           | 12%              | 8%              | 4.5 †                 | 32%           | 48%              | 20%             |
| 2.8 *                 | 0%            | 4%               | 96%             | 4.6 †                 | 52%           | 12%              | 36%             |
| 2.9 *                 | 0%            | 24%              | 76%             | 4.7 †                 | 36%           | 36%              | 28%             |
| 2.10 *                | 0%            | 12%              | 88%             | 4.8 †                 | 48%           | 20%              | 32%             |

**Key**

* In these items have low discrimination, almost everyone gets them right or almost everyone gets them wrong.
† In these items there is a broad spread in whether candidates address this issue; they discriminate well (<=60% are marked right or wrong in the item).
The high proportion of "not addressed" items in the clinical questions (questions 2 and 3) is of some concern. For example, items 2.2 (exchange pleasantries), 2.5 (enquire about progress of pregnancy), 2.6 (check notes), 3.2 (details of hysterectomy), 3.4 (types of incision), 3.5 (removal/conservation of ovaries in general), and 3.6 (oophorectomy if unexpected disease is found) are all very important aspects of everyday clinical practice, whether in terms of communication with patients or in obtaining preoperative consent. Failure to meet these standards may carry medical/medico-legal implications. Some candidates may have had a problem in this area. However, the candidates may not totally be to blame. The wording of the questions, the manner in which the questions were delivered and the wording of the marking schedules were potential influences, on the responses. These issues may have raised the proportion of "not addressed" questions. This could be addressed by exploring with candidates whether they felt that aspects of the examination had adversely influenced marking of their answers.

Table 4 provides a useful schedule against which to assess an examination. The 'purpose', 'aims' and 'stakes' involved in the examination are clear. Aspects of content validity are also addressed. However, five of the categories: consent to treatment and examination by medical students, clinical training, use of tissue and professional disagreement have only partially been addressed in the marking schedule. Several items may also be seen as evaluating clinical rather than ethical principles. These include: 2.6 'check notes for any relevant past medical/O&G his-

| CRITERIA               | ASSESSMENT                                                                                                                                 |
|------------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| 1. Purpose             | The test was primarily developed to reward excellence as high scoring candidates were appointed to SpR posts.                               |
| 2. Aim                 | The examination has both a retrospective and prospective focus.                                                                               |
| 3. Stakes involved     | The stakes were high as employment was involved.                                                                                             |
| 4. Content validity    | This was based upon the experience of NNA who drew up the examination paper on the basis of his professional experience and role as an organiser of Part II MRCOG courses. Content validity was also provided by following the RCOG Ethical Guidelines document. The questions were designed to address the whole domain covered in the RCOG document. Both the availability of relevant background documentation but conversely the restriction of available time reflect aspects of practice in real life. Candidates were provided with a copy of the guidelines so that the test had a clear relationship to the 'objectives of instruction'. The assessment had both a retrospective and a prospective focus as all the doctors had some experience of Obstetrics and Gynaecology but were joining a training scheme with a view to further training. |
| 5. Consequential validity | Long term effects of the undertaking the test were not assessed.                                                                        |
| 6. Construct validity  | This is the most difficult criteria to meet, as there is always the possibility of confounding factors, for example, communication skills, language skills, culture and gender affecting the results of an examination. Further work could be done to correlate this examination with other tests and to explore the thinking process of students in deciding the responses they made. |
| 7. Predictive validity | The long term performance of candidates who scored well in this test as opposed those who did not had not been assessed.                   |
| 8. Reliability         | Some evidence of test re-test reliability can be drawn from examining the difference of the means for the two groups of examinees examined on different occasions. Inter-rater reliability was assessed by the Kappa coefficient. Inter-rater reliability could be further assessed by examining the scores produced by a wider panel of assessors. Intra-rater reliability was assessed by the intra-class correlation coefficient. |
| 9. Generalizability    | The stability of the examination over time and in different contexts has not yet been addressed. This could also be addressed by checking the scores of the same group of candidates examined on different occasions and different groups of doctors in different hospitals. |
A modified OSCE assessing the assimilation and application of ethical principles relevant to obstetric and gynaecological practice.

In real life, doctors increasingly have access to documents, particularly via the internet and this is reflected in this OSCE. However, the effect of giving the RCOG document on “Professional Competence” beforehand is not clear. It may have influenced the marks achieved and this merits further investigation. Potential bias due to ethnicity has not been assessed in this study although it has been shown to have a small effect in a recent study. A fully acted out OSCE using actors could have been used, as opposed to a modified OSCE, as it has the advantage of mimicking real life more closely. However, it is also much more resource intensive. Further work could also be done in the areas of consequential, construct and predictive validity as there is always a question as to whether answers in examinations carry over into practice in real life. However, assessment of the modified OSCE against the schedule in Table 4 suggests it is of a reasonable standard.

Conclusion

This modified OSCE examination demonstrates the feasibility of testing ethical principles in Obstetrics and Gynaecology and on candidates for postgraduate posts and provides a basis for further work. It meets most of the criteria laid down in a checklist developed to assess postgraduate medical examinations. This modified OSCE can appropriately be used to assess the assimilation and application of a range of ethical principles applicable to Obstetric and Gynaecological practice. We have also provided tentative evidence that a modified OSCE may be an appropriate method for selecting candidates for postgraduate training schemes. Finally, our results suggest that a number of postgraduate doctors may have deficits in this important area of competency.

References

1. Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. Br Med J 1975;1(5955):447-51.

2. Tervo RC, Dimitrievich E, Trujillo AL, Whittle K, Redinius P, Wellman L. The Objective Structured Clinical Examination (OSCE) in the clinical clerkship: an overview. S D J Med. 1997 May; 50(5):153-6.

3. Royal College of Obstetrics and Gynaecology. Ethical Considerations relating to good practice in Obstetrics and Gynaecology. London: Available from http://www.rcog.org.uk/guidelines/ethics/ethical.html (Accessed 08.02.02)

4. Hutchinson L. Are medical postgraduate assessments valid? A systematic review of the published evidence. (MSc Thesis) Cardiff: University of Wales College of Medicine, 2000.

5. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professionals. Educational Researcher, 1995;24:5-11,35.

6. Wass V, Roberts C, Hoogenboom R, Jones R, van der Vleuten C. Effect of ethnicity on performance in a final objective structured clinical examination: qualitative and quantitative study. BMJ 2003; 326: 800-803

7. McFaul PB, Taylor DJ, Howie PW. The assessment of clinical competence in obstetrics and gynaecology in two medical schools by an objective structured clinical examination. Br J Obstet Gynaecol. 1993 Sep;100(9):842-6.

8. Descargues G, Sibert L, Lechevallier J, Weber J, Lemoine JP, Marpeau L. Evaluation of clinical competence in gynecology obstetrics: an innovative approach using the Objective Structured Clinical Examination]. J Gynecol Obstet Biol Reprod (Paris). 2001 May;30(3):257-64. French.

9. Dobay KJ, Nalesnik S. Lapar-OSCE: a laparoscopic observed structured clinical experience. Obstet Gynecol. 2001;97(4 Suppl 1):S8-S9.

10. Grand'Maison P, Blouin D, Briere D. Utilization of the objective structured clinical examination (OSCE) in gynecology/obstetrics. Proc Annu Conf Res Med Educ. 1985;24:65-70.

11. Singer PA, Robb A, Cohen R, Norman G, Turnbull J. Performance-based assessment of clinical ethics using an objective structured clinical examination. Acad Med. 1996;71(5):495-8.
van Woerden H, Agho F, Amso N.N, Stokes I. A modified OSCE assessing the assimilation and application of ethical principles relevant to obstetric and gynaecological practice.

12. Singer PA, Cohen R, Robb A, Rothman A. The ethics objective structured clinical examination. J Gen Intern Med 1993;8(1):23-8.

Correspondence

Dr H van Woerden
Specialist Registrar in Public Health Medicine
Temple of Peace and Health,
Cathays Park, Cardiff, CF10 3NW

Email address: hugo.vanwoerden@nphs.wales.nhs.uk
Fax: 02920402504.
Appendix 1

MARKING SCHEDULE

Marks Allocated 2 1 0

Question 1: What is your understanding of “Professional Competence?”
1.1 Appropriate training
1.2 Up to date with knowledge (Continued Professional Development)
1.3 Safe application of one’s surgical skills
1.4 Understands elements of good medical practice (attitude/ethical obligations)
1.5 Knowledge and application of clinical guidelines in own practice

Question 2: You and a medical student are about to see a 30-year old primigravida in the antenatal clinic. You have never seen her before. She is in her 36th week and the notes indicate that the pregnancy, so far, has been absolutely fine. She wishes to have an elective CS purely to avoid going through the pains of vaginal delivery. Explain your approach and the steps you take to achieve a satisfactory consultation.

2.1 Introduce oneself
2.2 Exchange pleasantries
2.3 Introduce medical student
2.4 Ask permission for his/her presence
2.5 Enquire about pregnancy progress and if any problem has arisen lately
2.6 Check the notes for any relevant past medical/O&G history
2.7 Ask if she has discussed the elective CS with her midwife, partner and GP
2.8 Explain the methods of pain relief available during labour
2.9 Explain the risks of CS, immediate (anaesthetic, bleeding, etc) and long term
2.10 If insisting on CS, either agree (women’s choice) or refer for a second opinion

Question 3: You are about to see a woman aged 40 years who is requesting hysterectomy for a confirmed diagnosis of DUB. Outline the main points of your discussion with her.

3.1 Alternative options (none, medical, Mirena, conservative surgery)
3.2 Details of hysterectomy & reference to total and subtotal
3.3 Risk of complications (bleeding, infection, ureteric injury)
3.4 Types of incision and routes of hysterectomy, type of anaesthesia
3.5 Removal and conservation of ovaries in general
3.6 Does she agree to oophorectomy if unexpected disease is found?
3.7 Thorough documentation of the above

Question 4: A medical device company invites you to the launch of its new device for the treatment of DUB. You are excited by its simplicity and low cost and want to use it in your hospital. Explain how will you achieve that.

4.1 Seek all available evidence on its safety and effectiveness
4.2 Where were the trials done and how long was the follow up period
4.3 Does it require GA, local or overnight stay
4.4 Can it be done in the outpatient or day surgery unit?
4.5 If good evidence exists, speak with clinical director & business manager
4.6 If no evidence is available, then should be introduced as a research project
4.7 Define safety and clinical outcome issues and cost
4.8 Determine resource implications for the health service

Key
2 = fully addressed, 1 = partially addressed, 0 = not addressed.