Applicability and fairness of the oral examination in undergraduate psychiatry training in South Africa

DJH Niehaus[1], E Jordaan[1], L Koen[1], M Mashile[1], S Mall[2]
[1]SELSUS Centre, Department of Psychiatry, University of Stellenbosch, South Africa
[1]Biostatics Unit, Medical Research Council, Bellville, South Africa
[2]Department of Psychology, University of Stellenbosch, South Africa

Abstract
Objective: There are several methods of evaluating medical students' performance, such as written examination, oral examination and objective structured clinical examination (OSCE). Many studies have focused on the reliability and validity of these methods but few studies have explored comparison between these methods. Psychiatry is the only subject at the University of Stellenbosch where the final assessment consists of solely an oral component. The aim of the study was to compare students' final overall and discipline specific examination marks (i.e. in the other subjects) with the examination marks in psychiatry, and to determine if content or structure (e,g. oral, written or OSCE format) of examination impacts more on the student performance in the examination. **Method:** 343 final year medical students were included. All undertook their psychiatry rotation at the University of Stellenbosch, South Africa during 2008 and 2009. Data of marks obtained in all the disciplines during 2008 and 2009 were collected and class marks were compared with their final examination marks across all disciplines. Bland-Altman plots were used to assess the level of agreement between the class and examination marks. Cases below the lower threshold were compared to all other cases across all disciplines. The odds ratio for group status was calculated for gender distribution of examiners. **Results:** The psychiatry class mark and final oral examination mark provided similar measures within a width of 31.5. Cases below the threshold had poorer performance in two other disciplines. The gender distribution of the examiners (female-female) significantly increased the odds ratio for poorer performance in the oral examination.**Conclusion:** The results suggest that a group of students underperform in their final examination independent of method of evaluation and that the gender of examiners impacts on examination marks. Therefore future research should focus on identifying and modifying factors (including choice of examiner combinations) that contribute to the poor performance of medical students in their final examination, in order to help students perform better. Gender distribution of examiners should also be considered when examinations are structured and designed.

Keywords: Oral; Medical; Student; Gender; Examiner; Performance; OSCE

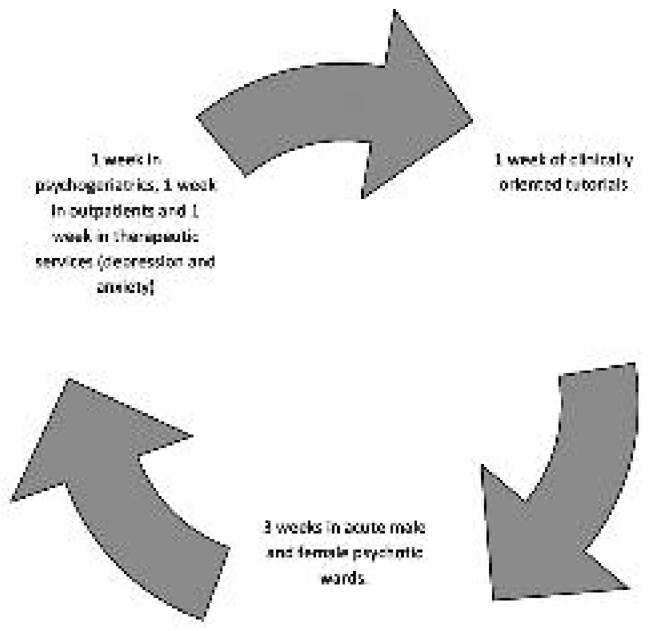
Received: 02-03-2011
Accepted: 23-06-2011
doi: http://dx.doi.org/10.4314/ajpsy.v15i2.16

Introduction

Within medical education throughout the world's universities, there are several means of evaluating student performance in examinations. These include written and oral examinations as well as objective structured clinical/practical examinations (OSCE/OSPE).[1,2] Health Sciences faculties in South Africa favour oral and OSCE/OSPE type examinations. Both of these modalities have specific advantages and disadvantages. OSC(P)E is a competence based examination that has been shown in many medical settings to assess clinical skills and knowledge in a reliable and valid fashion, a reflection of how communication skills are taught to medical students. In this method, learners are assessed by direct observation of their ability to communicate with simulated patients. The standardized evaluation setting in which this takes place reflects real life clinical encounters and the context of situations or problems that learners will encounter in actual medical practice. OSC(P)Es vary in length and scoring is done with a task specific checklist or a combination of checklists and rating scales. Studies exploring the link between OSCE and student performance have produced varying results. While some[3,4] have shown that OSC(P)Es have the ability to predict future postgraduate performance and that they motivate students to actively improve their clinicial skills others

Correspondence:
Ms S Mall
Department of Psychology, University of Stellenbosch
Private Bag X1, Matieland, 7602, South Africa
email: Sumaya.mall@gmail.com

Figure 1: Schematic representation of final year medical students' 7 week psychiatry rotation


1 week in psychogeriatrics, 1 week in outpatients and 1 week in therapeutic services (depression and anxiety)
1 week of clinically oriented tutorials
3 weeks in acute male and female psychotic wards.

have shown that they could induce anxiety and that the level of anxiety does not change significantly as students progress through the examination.[5,6]

The oral examination is a traditional form of assessment in which one or more examiners direct questions at the candidate. The aim of this type of examination is to assess factual knowledge and test other qualities such as mental agility. An advantage of oral examinations is that they have not been proven any more stressful than other forms of assessment.[7]

A number of disadvantages of this type of assessment have also been identified. These include perceived low reliability. The low reliability relates in part to the examiners active participation in the examination which can introduce bias.[7] As the oral examination is short (approximately twenty minutes), it is only possible to explore a small part of the field sufficiently. It is also possible that examiners can be biased or subjective.[8] Cox[9] suggests that orals can be highly threatening for candidates with resultant poor performance. Goldney and McFarlane[10] suggest that students who were successful in oral examination were more able to pick up cues and respond appropriately to the examiners. A concern often expressed by the examination boards is that some candidates have an advantage in oral examinations as a result of the way they present (communication ability) rather than what they present (medical skill and knowledge). Stupart et al[8] investigated differences in examination scores of medical students' clinical and surgery oral exams. Although they failed to find evidence of examiner bias, they did express concern of disparities in the marks of students from different racial backgrounds and between males and females. Marks of black, African students were considerably lower than their white peers and females generally tended to score higher than males. Lunz and Bashook[11] found mixed results for the influence of communication ability on candidate scores.

Currently psychiatry training of final year medical students at the University of Stellenbosch, South Africa, consists of a 7 week clinical rotation (including 1 week of clinically orientated tutorials) where students are divided into small subgroups rotating through various clinical areas: 3 weeks in the acute male and female psychotic wards and 1 week each in units for psychogeriatrics, outpatients and therapeutic services (depression and anxiety) (see Figure 1 for a schematic representation of the rotation). During the 7 week rotation students are trained in evaluating patients and making five axial diagnoses. They are also expected to interview patients under the guidance of a registrar or consultant, attend ward rounds and actively participate in the multi-disciplinary management of patients. Students' knowledge is assessed weekly by means of a computer-based test, the average of all the tests comprises 40% of their knowledge mark. On the last day of their rotation they write an OSPE which comprises 60% of their skills mark. Their final rotation (called class mark) mark (based on a combination of knowledge, skills and attitude) is calculated as follows: 60% OSPE and 40% knowledge/portfolio (includes peer assessment, case studies, ethics reviews and critical review of a relevant publication). The class mark constitutes 50% of the final exit mark, the other 50% made up by the psychiatry exit examination.

The psychiatry exit examination consists of one 20 minute oral examination with two examiners covering at least three diverse topics. Examiners are paired on the basis of language preferences (at least one of the two examiners must fully understand Afrikaans whilst both must be proficient in English). Students are allocated to examiners in alphabetical order. Failure in the first 20 minute oral leads to a second 20 minute oral with two different examiners on the same day. In contrast the other disciplines use mostly OSC(P)E structures. For example obstetrics and gynaecology utilize a 12 station OSCE. Seven of these stations are manned and 4 of these seven stations consist of standardized case discussion (7 minutes per station). The Ear-Nose and Throat division utilize a two station structure. One station comprises a practical evaluation of student skills whilst the second station consists of an oral examination with one examiner.

The aim of this study was to compare student's examination marks with their final mark, focusing specifically on psychiatry and determining whether students underperform in the oral examination, with the intention of identifying factors that may impact on their performance in the oral examination. To our knowledge, there is a paucity of research in South Africa comparing structure versus content of examinations and the ways in which these can impact on the performance of the student concerned. To our knowledge, there is also limited research on the impact of gender of examiners on student performance in the examination concerned. We believe that this study is the first of its kind to investigate the impact of gender of examiners on medical students' performance in the psychiatry examination.

Method

Study population

We recruited three hundred and forty three final year (sixth-year; 135 male and 208 female) medical students who presented to their final psychiatry rotation during 2008 and 2009. We then collected data on class marks, examination marks and final marks (examination plus class marks) obtained by students in all the disciplines.

Site

SELSUS Centre at Stikland Hospital. As part of the continued assessment during the final year medical students' psychiatry module, the SELSUS unit was created to structure, assess and research student teaching at Stikland psychiatric hospital. This centre includes a 24 hour SELSUS video surveillance area to enable observation of student-patient interviews and feedback of their performance.

Ethical considerations

The study was approved by the Committee for Human Research of the Faculty of Health Sciences at the University of Stellenbosch, and permission was granted to access marks only once the students had completed their final year and all information were regarded as confidential. We then collected data of marks obtained by students in all the disciplines during year 2008 and 2009.

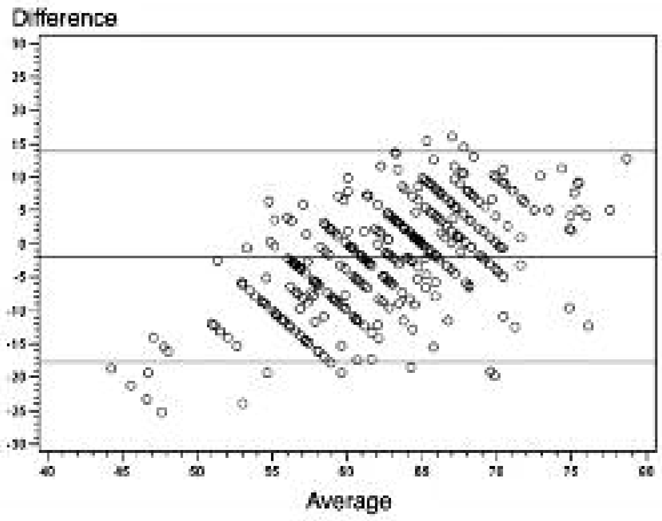
Analysis

A descriptive analysis of the exit examination marks for each student and the group as a whole was performed using SPSS version 16.0 (Statistical software package for social scientists). Bland-Altman plots were used to assess the level of agreement between the class mark and the examination mark. Cases below the lower threshold were compared to all other cases across all disciplines and the odds ratio for group status was calculated for gender distribution of examiners and students.

Results

The students' class and examination marks are summarized in Table I. The mean overall examination mark for students was 63.4% (range 48.7-78.4, SD 5.66). The mean psychiatry exit examination mark (oral) was 61.78% (range 35.0-85, SD 9.03) and showed no significant difference from the class mark (mean 63.70%, range 51.64-82.27, SD 5.17). The Bland Altman plot showed a mean difference of -1.93 (95% limits of agreement = -17.66 to 13.80) and a width of 31.5. The limits of agreement are only estimates of the values which apply to the whole population and we used standard errors and confidence intervals to give an indication of the precision of our estimates. Given a 95% confidence interval for the bias of - 2.79 to -1.07, for the lower limit of agreement of -19.15 to -16.17 and for the upper limit of 12.32 to 15.29, the most optimistic interpretation is that there can be a difference of 12 points between the two evaluation methods (Figure 2).

Figure 2: Bland Altman analysis of psychiatry class and oral marks. Difference against mean for psychiatry class and examination marks



Mean difference (bias) = -1.93% ; 95% limits of agreement-17.7% to 13.8%

Table I: Psychiatry class mark and examination marks

Discipline		*Minimum*	*Maximum*	*Mean*	*Std. Deviation*
Anaesthesiology	CM	50.0	85.0	71.1	5.8
	EM	36.0	91.0	62.2	10.5
Ear, Nose & Throat	CM	42.0	85.0	66.5	8.4
	EM	45.0	80.0	63.4	7.3
Family Medicine	CM	60.3	77.1	70.0	3.1
	EM	40.3	79.0	62.7	6.2
Internal Medicine	CM	57.4	77.4	66.5	3.4
	EM	40.0	81.0	62.9	7.1
Obstetrics & Gynaecology	CM	49.7	81.9	66.2	6.0
	EM	40.0	87.0	62.9	8.3
Orthopaedics	CM	43.0	81.0	65.2	6.3
	EM	40.0	85.0	62.3	8.6
Paediatrics	CM	54.5	78.5	66.2	4.3
	EM	42.0	80.0	61.4	6.8
Psychiatry	CM	51.6	82.3	63.7	5.2
	EM	35.0	85.0	61.8	9.0
Surgery	CM	51.5	79.9	67.1	5.0
	EM	28.0	81.0	63.3	7.0
Urology	CM	52.0	82.0	68.5	5.9
	EM	42.0	80.0	64.3	7.6

CM=class mark, EM=examination mark

Table II: Bland Altman statistics across disciplines

Subject	*Bias (95% CI)*	*95% Limits of agreement*	*Width*
Family Medicine	-4.5 (-5.0 to -3.96)	-14.4 to 5.4	19.8
Internal Medicine	-3.7 (-4.2 to -3.1)	-14.4 to 7.0	21.4
Surgery	-3.8 (-4.4 to -3.4)	--14.4 to 7.1	21.8
Pediatrics	-4.75 (-5.4 to -4.1)	-16.5 to 7.0	23.5
Obstetrics and Gynecology	-3.33 (-4.0 to -2.7)	-15.7 to 9.1	24.8
Urology	-4.12 (-4.8 to -3.4)	-17.1 to 8.9	25.9
Psychiatry	-1.93 (-2.8 to -1.11)	-17.7 to 13.8	31.5
Ear. Nose and Throat	-3.11 (-4.0 to -2.2)	-19.5 to 13.3	32.8
Opthalmology	-0.89 (-1.9 to – 0.13)	-19.5 to 17.7	37.2
Anesthesiology	-8.94 (-10 to -7.9)	-28.3 to 10.5	38.9
Overall mark	-2.88 (-3.8 to -2.0)	-19.1 to 13.3	32.4

We defined a threshold for agreement equal to the lower limit of agreement (-17.66) and identified the students belonging to this group (n=12). This group did at least 17.7 points better in their class mark than in examination (see Tables II and III). The exam performance of this group (12 students) was then compared to that of the other students for all subjects but psychiatry. The average exam mark for all 12 subjects did not differ between the 2 ''performance'' categories for psychiatry (p=0.3861) but the group did perform significantly worse in the obstetrics (58% vs 63%) and ear, nose and throat (58% vs 63%) examinations compared to their class mark in each of these subjects.

The gender of a student was not a significant predictor for being part of the group below the threshold (p=0.676), but the gender combination (MF, MM, FF) of the examiners was a significant predictor (p=0.029).The majority of the students (59.3%) were assessed by a male-female examiner combination, whilst 28.9% were seen by the male-male examiner group and 11.8% by the female-female examiner combination group. Odds ratio estimates showed that compared to a male/female examiner combination, a male/male combination did not increase the chance of a below threshold mark (OR=3.5; 95% CI 0.8 to 15.2). The odds of scoring below the threshold is 7.4 (1.6 to 34.4) times higher when the examiner pair is female-female in comparison to when the examiner pair is male-female.

Table III: Group of students that scored below the threshold (-17.7) of agreement

	Examination mark (%)	*Class mark (%)*	*Difference between marks (%)*
1	60	80	-19.7
2	35	58	-23.3
3	37	56	-19.4
4	50	69	-19.3
5	35	60	-25.2
6	45	64	-19.4
7	41	65	-24.0
8	55	73	-18.5
9	60	79	-19.2
10	35	56	-21.2
11	50	68	-17.8
12	35	54	-18.6

Discussion

For the 343 medical students no significant difference could be detected in the overall performance of students in the above and below threshold groups defined by the disagreement in their psychiatry marks. On average, the class marks are higher than the exam marks.

However, when considering only the 295 students who passed the exam (above 50%), the bias is almost zero. The 95% limits of agreement indicate a large possible discrepancy between the exam and class marks for psychiatry. However, the limits compare favorably to that of the other subjects, indicating that the agreement between the exam and class marks for some of the other subjects were poorer than for psychiatry. The psychiatry class and oral exit examination marks provided similar measures assuming that examiners accept that the class mark may be almost 18 percentage points above the exam mark or 14 percentage points below the exam mark.

Indeed, the group of students that scored below the threshold (Bland Altman analysis) tended to underperform in at least two other disciplines that use different examination structures than that of psychiatry. This suggests independence from method of evaluation in students that underperform in their final examinations since in other disciplines the exam consists of different components of which oral exam forms only a part.

These results do not support the common belief that oral examinations should be avoided due to the perceived low reliability that is hypothesized to be linked to increased anxiety in oral settings, selective testing of knowledge (smaller area of work covered) and possible examiner bias.

We encountered limitations in the study. Firstly, our study did not assess anxiety state at the time of examinations and we cannot comment on the role of anxiety in the below threshold group. However, literature on test anxiety shows conflicting results with some studies showing a negative linear relationship between anxiety and performance on objective structured clinical examination[5], whilst some show similar anxiety levels across different methods of examination.[12] The latter study assessed performance of 3rd year medical students at University of Miami School of Medicine in oral examinations, computer generated written examinations and other behavioral evaluations. Their aim was to determine which students had debilitating anxiety associated with oral examinations and whether their performance in the oral examination differed significantly from performance in the other methods of evaluation. They reported that students with high anxiety who scored lower in the oral examination also had similar results (scored lowest of all groups) in the other methods of evaluation. Similarly, Retequiz[13] reported that both the standardized patient exam and multiple choice questions used in medical clerkship evoked similar levels of subjective anxiety, but that this anxiety did not influence exam performance. This suggests that anxiety related to oral examination is likely to be no different to anxiety experienced in other evaluations.

Oral examinations are believed to test only a small portion of the overall knowledge base. In our setting each oral examination covers at least three distinct topics and includes two examiners that each has 10 minutes to question the candidate. This structure could possibly limit the impact of the perceived lack of adequate coverage of a range of topics and avoid inappropriate in-depth coverage of single topics in the oral examination. This strategy by no means eliminates the critique on coverage. However, it could be argued that other forms of examination (written, OSC(P)E) may also be guilty of only partially testing some of the educational domains (knowledge, comprehension, application, analysis, synthesis and evaluation).[14] The success of any of these methods may be hypothesized to rely on the knowledge, experience and interpersonal skills of the examiners and examinees, rather than on the method itself.

Examiner bias is a more difficult aspect to measure.[8] Reasons for differences between examiners may include ethnicity, age, gender, personality and such practical issues as their manner with students, their ability to construct/convey the question in a more or less understandable way than another examiner or their competency level in assessing the answer of the student.[15] We ascertained the gender distribution of the examiners in an attempt to address gender-bias in the oral setting. The gender of the examiners did impact on the outcome of oral examination and a female/female combination increased the chances of a student to underperform in the oral examination. This finding suggests that the choice of examiner pairs should be based on gender and that conveners of oral examination should preferably utilize male/female combinations. The reason for this gender effect is not clear and ongoing studies at our site are focusing on the possible role of non-verbal behaviour cues and gender in the examination setting. These findings, based on a small sample of 12 students, should be interpreted cautiously as our study did not include an objective measure of anxiety and other unknown contributing factors to performance in oral exams that may confound the findings.

Conclusion

The aim was to determine whether students underperform in the oral examination versus other forms of examination, and the results showed that students who underperformed in the psychiatry oral examination also underperformed in other components of their final examination, suggesting an independence from method of evaluation. This conclusion is however subject to the assumption that the width of 31.5 is acceptable for examiners. The results of this study should not be seen as supporting the oral examinations as the ‘‘gold standard’’, but should rather be seen as a challenge to re-assess the role of oral examination in disciplines where interpersonal communication and synthesis of knowledge are critical to success in the workplace. One should also be asking what factors, in addition to gender combination of examiners, contribute to poor performance of medical students in their final examination. Research should focus on identifying and modifying these factors.

References

1. *Jeffries A, Roukema H, Skimore M, Herol J. Using an OSCE to assess multiple physician competencies in postgraduate training. Medical Teacher 2007; 29(2-3):183-191.*
2. *Smith LJ, Price DA, Houston IB. Objective structured clinical examination compared with other forms of student assessment. Arch Dis Child 1984 Dec;59(12):1173-1176.*
3. *Wallenstein J, Heron S, Santen S, Shayne P, Ander D. A core competency-based objective structured clinical examination (OSCE) can predict future resident performance. Acad Emerg Med 2010 Oct;17 Suppl 2:S67-71.*
4. *Gupta P, Dewan P, Singh T. Objective Structured Clinical Examination (OSCE) Revisited. Indian Pediatr 2010 Nov 7;47(11):911-920.*
5. *Brand HS, Schoonheim-Klein M. Is the OSCE more stressful? Examination anxiety and its consequences in different assessment methods in dental education. Eur J Dent Educ 2009 Aug;13(3):147-153.*
6. *Schiff R. A short case prolonged. BMJ: British Medical Journal (International Edition) 2001 09/08;323(7312):551.*
7. *Kelley PR, J., Matthews JH, Schumacher CF. Analysis of the oral examination of the American Board of Anesthesiology. J Med Educ 1971 11;46(11):982-988.*
8. *Stupart D, Goldberg P, Krige J, Khan D. Does examiner bias in undergraduate oral and clinical surgery examinations occur? S Afr Med J 2008 Oct;98(10):805-807.*
9. *Cox KR, Ewan CE. The Medical teacher. Edinburgh : New York: Churchill Livingstone; 1982.*
10. *Goldney RD, McFarlane AC. Assessment in undergraduate psychiatric education. Med Educ 1986 Mar;20(2):117-122.*
11. *Lunz ME, Bashook PG. Relationship between candidate communication ability and oral certification examination scores. Med Educ 2008 Dec;42(12):1227-1233.*
12. *Linn BS, Zeppa R. Anxiety and performance on oral examination. Soc Sci Med E 1981 Aug;15(3):211-214.*
13. *Reteguiz JA. Relationship between anxiety and standardized patient test performance in the medicine clerkship. J Gen Intern Med 2006 May;21(5):415-418.*
14. *Slavin RE. Educational psychology: theory and practice. 7th ed. Boston: Allyn and Bacon; 2003. p. 466-467.*
15. *Vickers M. Quality Assurance in Postgraduate Examinations. Baillieres Clinical Anaesthesiology 1994 SEP;8(3):711-726.*