

Paper versus electronic feedback in high stakes assessment

AJ Munro¹, K Cumming², J Cleland³, AR Denison⁴, GP Currie⁵



Tablet computers have emerged as increasingly useful tools in medical education, particularly for assessment. However, it is not fully established whether tablet computers influence the quality and/or quantity of feedback provided in high stakes assessments. It is also unclear how electronically-recorded feedback relates to student performance. Our primary aim was to determine whether differences existed in feedback depending on the tool used to record it.

Methods We compared quantitative and qualitative feedback between paper-scoring sheets versus iPads™ across two consecutive years of a final year MBChB (UK medical degree) Objective Structured Clinical Examination. Quality of comments (using a validated five-point rating scale), number of examiner comments and number of words were compared across both methods of recording assessment performance using chi-squared analysis and independent t-test. We also explored relationships between student performance (checklist and global scoring) and feedback.

Results Data from 190 students (2850 paper scored interactions) in 2015 and 193 (2895 iPad™ scored interactions) in 2016 were analysed. Overall, a greater number of comments were given with iPad™ compared to written (42% versus 20%; $p < 0.001$) but the quality of feedback did not differ significantly. For both written and electronic feedback, students with low global scores were more likely to receive comments ($p < 0.001$).

Conclusion The use of iPads™ in high stakes assessment increases the quantity of feedback compared to traditional paper scoring sheets. The quantity and quality of feedback for poorer performing candidates (by global score) were also better with iPad™ feedback.

Keywords: assessment, feedback, objective structured clinical examination, tablet computers

Declaration of interests: No conflict of interests declared

Introduction

Assessment plays an important role throughout all stages of both under- and postgraduate medical education. As well as determining that a certain standard has been reached, ranking students, highlighting areas of weakness in teaching or the curriculum, the assessment process can serve as a means by which feedback can be given to students and demonstrate areas where knowledge, skills and attitudes need improved.^{1,2} The Objective Structured Clinical Examination (OSCE) is a well-established tool in the armamentarium of assessors and simple binary checklists are often employed to record marks.³ OSCEs also lend themselves to examiners and patient partners being able to record global rating scales according to overall performance, while written comments have the potential

to provide a wealth of information, especially in poorly performing or 'borderline' students.^{4,5} However, issues can arise when checklist paper marks recorded are 'read' by optical software and missing marks are found.⁶ Moreover, written comments about performance (typically encouraged in less well performing students) may not have a dedicated space and be considered as an afterthought by examiners, hand-writing may be difficult to read and providing students with a record of such individualised comments may prove difficult. This has led to interest in, and increasing use of, mobile electronic tools in an attempt to streamline and enhance different aspects of the assessment process.

Studies have looked at the reliability of electronic marking efficiency and the process and acceptability to examiners and candidates.^{7–9} While useful at progressing knowledge, it

Correspondence to:

G Currie
Institute of Education
for Medical and Dental
Sciences
School of Medicine,
Dentistry and Nutrition
University of Aberdeen
Aberdeen AB25 2ZD
UK

Email:

graeme.currie@nhs.net

¹GPST1, Banchory Group Practice, Banchory, Aberdeenshire; ²Clinical Fellow, Emergency Department, Monklands Hospital, Lanarkshire; ³John Simpson Chair of Medical Education Research, ⁴Professor of Medical Education, School of Medicine, Dentistry and Nutrition, University of Aberdeen, Foresterhill, Aberdeen; ⁵Consultant Chest Physician, Aberdeen Royal Infirmary, Aberdeen, UK

is also important to look beyond the basics, and determine if the use of electronic devices leads to improved feedback. In one of our early studies, it was found that the use of iPads™ in a Year 1 summative OSCE resulted in an increase in quantity and quality of individual comments compared to paper-based checklists.⁴ However, it is uncertain if these findings can be extrapolated to higher stakes examinations, which are typically longer, and of greater complexity. It is also possible that examiner behaviour in providing feedback may vary in assessments of different degrees of importance.

Feedback is a vital component of medical education and is necessary for students and trainees to improve as learners and clinicians.^{10,11} It should be focused, given immediately, help direct future learning, and be a vehicle by which to reinforce correct behaviours and competencies.^{12,13} In this respect, individualised feedback (rather than binary checklist items) based on a students' performance during an OSCE can often be a rich source of information, especially in less well performing students where explicit feedback may confer greatest benefit.

Our primary aim was to determine if the quality and quantity of examiner written feedback relating to candidate performance in a Final Year MBChB (the qualification that students typically obtain from UK medical schools in order to practise medicine) OSCE differed depending on whether feedback was recorded by hand or electronically. Our secondary aim was to explore the relationship between comments and performance.

Methods

We conducted a retrospective database and exam sheet analysis exploring quantitative and qualitative feedback across two methods of capturing data. Comparisons were made between traditional paper scoring sheets versus iPads™ across two consecutive years (2015 and 2016) of a final year student MBChB 15 station OSCE at the University of Aberdeen.

In 2011, the University of Aberdeen designed an electronic OSCE application (app) for use on iPads™, containing all the components of the paper-based marking sheet including a free space section for optional written comments. As well as marking a box to record completion of checklist items, examiners are asked to document a global score for each encounter, both of which contribute to the production of a pass mark for that particular station using the borderline regression method.¹⁴ The global score used consisted of a 1–5 Likert scale where 1 indicated an 'unsatisfactory' and 5 indicated an 'excellent' performance, and is considered to be an important reflection of students' overall ability.¹⁵ At pre-exam briefings, all examiners were encouraged to provide written (in 2015) or electronic feedback (in 2016) where considered appropriate, especially in poorly performing and 'borderline' students. No specific training was provided on how to give 'good' feedback. All examiners were aware their comments could be provided to the students.

For each student–examiner encounter, data were collected depending on whether an examiner comment was made and, if so, the number of words per comment. The global rating per station was also collected. Comments from both years were scored independently for quality by two authors (AJM and KC) using a validated 5-point rating scale (Feedback Quality Rating Scale; Kappa 0.625 for inter-rater reliability).⁴ Quality point descriptors were as follows:

1. judgmental non-specific praise or comments on appearance only (lowest quality of feedback)
2. description of performance or suggestion for improvement
3. description of performance *and* suggestion for improvement
4. objective appraisal of performance
5. objective appraisal of performance *and* suggestion for improvement (highest quality of feedback).

Where discrepancies occurred between quality scores for individual comments, they were reviewed by both authors together to determine the most appropriate quality score. Such discrepancies were found in only 57 (3%) of all comments.

Data were stored and analysed using SPSS version 23.0. We compared the number of examiner comments, total number of words, mean number of words per comment and mean number of words per station for each OSCE station using independent t-test. Relationships between written and electronic feedback according to overall student performance were explored according to checklist and global rating scores provided by examiners, using chi-squared analysis.¹⁵ The study was reviewed by the local College Ethics Review Board and ethical permission was granted.

Results

Data from 190 students (comprising 2850 exam papers using paper based marking) in 2015 and 193 (comprising 2895 electronic data entries using iPad™ marking) in 2016 were analysed. Table 1 illustrates data relating to characteristics of comments. A significantly greater overall ($p < 0.001$) number of comments, total number of words and mean number of words per station were given with iPad™ versus written feedback. The mean number of words per comment was greater ($p < 0.001$) for written versus electronic marking (15 versus 12 respectively), indicating that written comments (when given), were of greater length than electronic comments. For both written and iPad™ feedback, most comments were quality rating 2 (i.e. second lowest quality of feedback; 76% versus 80% respectively) with no difference in overall quality ($p = 0.223$).

For all global ratings, students were significantly ($p < 0.001$) more likely to receive feedback when iPads™ were used for marking compared to paper checklists (Figure 1). For both written and electronic feedback, the most poorly performing students (according to the global rating scores) were significantly more likely to receive a high quality comment

Table 1 Comparison between paper and electronic feedback. Quality ranked 1–5 using the Feedback Quality Rating Scale, where 1 is the lowest and 5 is the highest quality of feedback

	Paper (2015)	iPad™ (2016)	p-value for the difference
Overall number of comments n (%)	548 (20)	1226 (42)	p < 0.001
Total number of words	8015	15040	p < 0.001
Mean number of words per comment (SD)	15 (11.0)	12 (8.4)	p < 0.001
Mean number of words per station (SD)	2.9 (7.7)	5.2 (8.2)	p < 0.001
Quality of comments, n (%)	1	32 (5.8)	p = 0.223
	2	417 (76.1)	
	3	9 (1.6)	
	4	86 (15.7)	
	5	4 (0.7)	

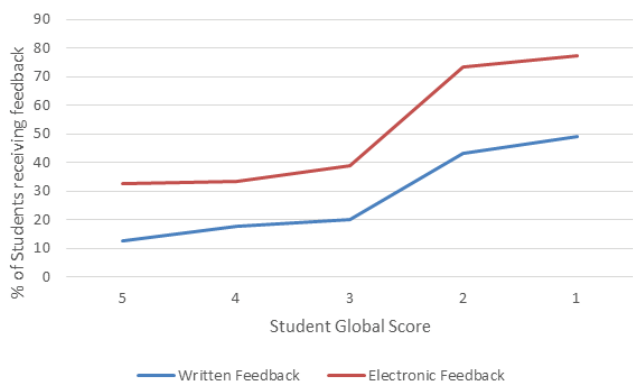


Figure 1 Proportion of students receiving feedback depending on modality of recording during assessment. For both modalities, students were significantly more likely to receive feedback if they performed poorly (p < 0.001). For each global score, a significantly (p < 0.001) greater proportion of students received feedback with use of an iPad

(as indicated by the Feedback Quality Rating Scale) than highly performing students (p < 0.001) (data not shown). For both marking modalities, the highest proportion of students receiving feedback was ‘unsatisfactory’ (i.e. a score of 1 on the global rating scale). For these students, comments were left for 49% for written versus 77% for electronic marking (p < 0.001 for the difference).

Table 2 demonstrates that for both written and electronic feedback, the global rating score had a significant impact on mean number of words of feedback provided (p < 0.001). In other words, poorly performing students received longer comments than highly performing students irrespective of mode of recording.

For both written and electronic feedback, students with lower checklist scores were significantly more likely to receive feedback than students with high checklist score (p < 0.001). As shown in Figure 2, for written feedback, the quality of feedback was also likely to be significantly greater for students with low checklist scores (p = 0.007). Checklist score did not, however, have a significant impact on the quality of feedback (as indicated by the Feedback Quality Rating Scale) given for electronic feedback (p = 0.882).

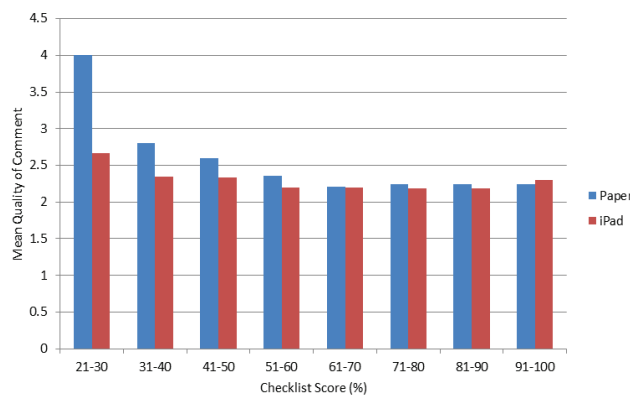
Discussion

Although tablet computers are acceptable, cost-efficient and result in fewer missed marks when used during assessment, it remains important to demonstrate they do not adversely influence quantity or quality of feedback.¹⁶ Our study found that the use of iPads™ in high stakes assessment resulted in a significantly greater quantity of feedback compared to written comments, without a negative impact on quality of comments. However, while overall feedback was not of a high quality, examiners provided significantly more, and better quality, comments for less capable students when using a tablet computer.

Our findings are in keeping with results from an earlier study in which an increase in the quantity of electronic comments versus paper was observed.⁴ In contrast to our own findings, in this earlier study, iPads™ did lead to a significant (p < 0.001) increase in the overall quality of comments. Why this was the case is unclear, although the authors postulated that the relative novelty of the iPad™ may have ‘encouraged’ better comments. In addition, there may have been fewer time constraints in a first year OSCE compared to a final year, allowing higher quality comments to be left. For early year students, action can be taken according to the nature of feedback whereas by final year, concerns may be difficult to remediate. This all means that poorly performing students

Table 2 Mean number of words of feedback provided in written and electronic form, according to global score

Global Score	Mean number of words of feedback	
	Written	Electronic
5 'Excellent'	11	11
4 'Highly satisfactory'	12	10
3 'Satisfactory'	13	12
2 'Borderline'	18	15
1 'Unsatisfactory'	23	17

Figure 2 Impact of checklist score on the quality of comments according to modality of recording during assessment. For paper feedback, checklist score had a significant impact on the quality of feedback ($p = 0.007$). Checklist score did not have a significant impact on quality of electronic feedback ($p = 0.882$)

can be 'signposted' to specific areas of weakness in their performance, which could be especially useful in formative assessment.

It is uncertain why there was an increase in the quantity of comments when iPads™ were used during assessment. However, iPads™ had a dedicated 'free type' section for optional comments which examiners were directed to at the end of the checklist, while paper marking sheets did not have a dedicated area for comments (although examiners were encouraged to write on the reverse side of the checklist). It is also possible that examiners found it quicker to type comments, given tablet and keyboard use are now ubiquitous in many work and recreational contexts. It is also possible that examiners found marking easier (and hence quicker) using the iPad™ and, as a consequence, had more time available to type comments.


The increase in quantity of provided comments was greatest in the group comprising of 'borderline' and 'unsatisfactory' students according to global score. These groups were also more likely to receive better quality comments. This is both reassuring and important, as these students are most likely to gain greatest benefit from individual feedback. There was not, however, an increase in quality for poorer performing students by checklist score. A further administrative benefit of tablet recording of comments is that they can be provided to all students, removing a requirement for written comments to be transcribed. This could be especially useful, as it

has been demonstrated that recall after receiving verbal feedback following an OSCE can be inaccurate and poorly representative.¹⁷

Our study has some important strengths. First, for both written and electronic comments, all were scored by two authors independently and a joint decision was made when there were discrepancies (typically due to illegible examiner handwriting). Second, the quality of comments was graded using a previously validated scoring system rather than individual impressions. Third, our study involved a large number of students and experienced examiners across 15 OSCE stations, increasing the chances of our findings being 'real' and not occurring by chance. In addition, the study was based in a real exam setting so feasibility and applicability are high.

While our paper adds to the few published reports on the topic, it has some weaknesses. For instance, while the examiner cohort, number and characteristics of students and OSCE station content/format were similar across both years, they were not identical. However, the cohort of examiners is similar from year to year and the exam itself was comparable, with the same organisers and a similar (low) pass mark. It is important to point out that our study was of a pragmatic observational nature with no controls, meaning there could be other reasons for our findings such as increased comfort with typing rather than hand-writing and greater number of prompts/cues. In this respect, an analysis of frequency and quality of typed comments in future final year OSCEs would be useful. Finally, the authors were not blinded to whether comments were written or electronic.

Further work is required to explore why the improvement in quality of comments observed in early year OSCEs was not replicated in a Final Year OSCE. In addition, the existing app could be developed further to prompt or mandate examiners to leave a comment if a 'borderline' or 'unsatisfactory' global score is recorded.

Our study has shown that the use of tablet computers in a high stakes OSCE can result in an increase in the quantity of bespoke comments relating to individuals' performance generally, and both the quantity and quality of comments/feedback for less well performing students was also better. These findings add to the literature supporting the use of tablet computers in assessment. Ongoing consideration should be given to ensuring that examiner training promotes best practice in improving the delivery of high quality feedback. 

Acknowledgements

The authors would like to thank Shona Fielding from the University of Aberdeen Medical Statistics Team for her assistance with the statistical analysis for this study.

References

- 1 Black NM, Harden RM. Providing feedback to students on clinical skills by using the Objective Structured Clinical Examination. *Med Educ* 1986; 20: 48–52.
- 2 Turner JL, Dankoski ME. Objective structured clinical exams: a critical review. *Fam Med* 2008; 40: 574–8.
- 3 Harden R, Stevenson M, Downie WW et al. Assessment of clinical competence using objective structured examination. *BMJ* 1975; 1: 447–51.
- 4 Denison A, Bate E, Thompson J. Tablet versus paper marking in assessment: feedback matters. *Perspect Med Educ* 2016; 5: 108–13.
- 5 Van Nuland M, Van den Noortgate W, van der Vleuten C et al. Optimizing the utility of communication OSCEs: omit station-specific checklists and provide students with narrative feedback. *Patient Educ Couns* 2012; 88: 106–12.
- 6 Snodgrass SJ, Ashby SE, Onyango L et al. Electronic practical skills assessments in the health professions: a review. *Internet J Allied Health Sci Pract* 2014; 12: 1–10.
- 7 Treadwell I. The usability of personal digital assistants (PDAs) for assessment of practical performance. *Med Educ* 2006; 40: 855–61.
- 8 Hochlehnert A, Schultz JH, Möltner A et al. Electronic acquisition of OSCE performance using tablets. *GMS Z Med Ausbild* 2015; 32: Doc41.
- 9 Schmitz FM, Zimmermann PG, Gaunt K et al. Electronic rating of objective structured clinical examinations: mobile digital forms beat paper and pencil checklists in a comparative study. In: Holzinger A., Simonic KM. (eds) *Information Quality in e-Health*. USAB. *Lecture Notes in Computer Science* 2011; 7058: 501–12.
- 10 Chowdhury RR, Kalu G. Learning to give feedback in medical education. *The Obstetrician and Gynaecologist* 2004; 6: 243–7.
- 11 Brukner H. Giving effective feedback to medical students: a workshop for faculty and house staff. *Med Teach* 1999; 21: 161–5.
- 12 Ende J. Feedback in clinical medical education. *JAMA* 1983; 250: 777–81.
- 13 van de Ridder JMM, Stokking KM, McGaghie WC et al. What is feedback in clinical education? *Med Educ* 2008; 42: 189–97.
- 14 Hejri SM, Jalili M, Muijtjens AMM et al. Assessing the reliability of the borderline regression method as a standard setting procedure for objective structured clinical examination. *J Res Med Sci* 2013; 18: 887–91.
- 15 Regehr G, MacRae H, Reznick RK et al. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998; 73: 993–7.
- 16 Currie GP, Sinha S, Thomson F et al. Tablet computers in assessing performance in a high stakes exam: opinion matters. *J R Coll Physicians Edinb* 2017; 47: 164–7.
- 17 Humphrey-Murto S, Mihok M, Pugh D et al. Feedback in the OSCE: what do residents remember? *Teach Learn Med* 2016; 28: 52–60.