Clinician and simulated patient scoring – the psychometrics of a national programme recruiting dentists to DF1 training posts

C. M. Wiskin,*[1] K. Elley,[2] E. Jones[3] and J. Duffy[4]

IN BRIEF

- Offers transparency about how the new national recruitment process is managed and validated.
- Provides information about how communication is scored (this field is generic to all specialties).
- Increases understanding of the use of simulated patients.
- Thinking about the importance of communication is relevant to anyone in a people-facing position.

EDUCATION

Introduction In 2012 a national, standardised approach was taken to UK Dental Foundation 1 recruitment. Prior to that recruitment method was at the discretion of individual Deaneries. The new national system is interactive, including simulated patients to see how applicants perform in a clinical communication context. A question was whether simulated patient scores could/should be awarded as well as clinicians' scores. This paper presents score data collected in the first round of national DF1 recruitment centres, with focus on how clinical examiners and trained simulated patients rated applicants. **Method** At the live recruitment events across four national centres score data were collected from observing clinical assessors and simulated patients on the communication station. On this occasion only the clinician awarded scores 'counted', but all simulated patients completed marking sheets to enable the process to be evaluated. Data were retrospectively analysed to test the hypotheses that there would be no significant scoring differences between centres and that inter-rater reliability, by applicant, between paired clinicians, and between clinicians and simulated patients would be strong. **Results** Results showed encouraging consistency between assessors, with some differences between centres. Clinicians were more likely to offer a borderline score. In communication analyses empathy had the weakest correlation with the overall score, while professional attitude had the strongest correlation. Data supported the hypothesis that trained simulated patients can be considered as assessors. Their future inclusion offers candidates a dual perspective (clinical and non-clinical) on performance, and saves clinical time. **Discussion** Simulated patients scored consistently and value can be added by including different perspectives in interactive assessment. Robust training is needed in all assessor training. **Conclusion** Simulated patients can usefully contribute to scoring in national dental recruitment centres. Lessons learned here can inform other dental assessments where stakeholders are already using, or considering using, simulated patients as assessors or co-assessors.

INTRODUCTION

This paper addresses the question of scoring reliability and consistency when recruiting applicants nationally to Dental Foundation 1 (DF1) posts using simulated patient stations. UK data are presented that shed light on questions about simulated patient inclusion as summative assessors, clinician inter-rater reliability, and sameness/difference between geographically diverse centres.

[1]*University of Birmingham, College of Medical and Dental Sciences, Edgbaston, Birmingham, B15 2TT; [2]Dean of Dentistry, West Midlands Postgraduate Deanery, St Chad's Court, 213 Hagley Road, Birmingham, B16 9RG; [3]Dean of Dentistry, London Deanary, Stewart House, 32 Russell Square, London, WC1B 5DN; [4]Primary Care Clinical Sciences, College of Medical & Dental Sciences, University of Birmingham, Edgbaston, B15 2TT
*Correspondence to: Dr Connie Wiskin
Email: c.m.wiskin@bham.ac.uk

Refereed Paper
Accepted 10 May 2013
DOI: 10.1038/sj.bdj.2013.739
©British Dental Journal 2013; 215: 125–150

UK DF1 training is the first year post qualification where new graduates have supervised experience working in primary dental care as part of the National Health Service (NHS) general dental service (GDS). This is NHS funded. The trainers are experienced GDS practitioners selected by Postgraduate Deaneries. This training is required for all new UK graduates wanting to work in the NHS. For overseas nationals there is an alternative route of foundation training by competence assessment, based on previous primary care experience. EU dentists are currently not required to undertake DF1 training, although some choose to apply. Although the numbers of DF1 training posts broadly match the numbers of UK graduates, applications by eligible others means demand for posts exceeds supply. Competition therefore exists for places, making this a high-stakes recruitment exercise. (The COPDEND website [UK Committee of Postgraduate Dental Deans and Directors www.copdend.org] is a useful resource for documents/reports relating to DF1 and other UK postgraduate dental training).

In 2012 a national (initially England and Wales, to include Northern Ireland for the next cycle), standardised approach was taken to DF1 recruitment for start posts for the first time. Prior to that recruitment strategies were localised, and at the discretion of individual Deaneries. Methods included paper only applications and paper plus panel interviews (the latter of which – in spite of good face validity – have reported reliability challenges[1]). This contrasted with the nationally adopted interactive vocational training scheme (VTS) centralised in 2004 for medical general practice. The VTS has been extensively researched and evaluated. Documents including the equalities impact report, history, evidence base, related research articles and practical guides are centrally available.[2]

Dentistry recognised the advantages of the VTS model, and the need to have a single process nationally, leading to the formation of the National DF1 Recruitment Working Group in spring 2011.

At the West Midlands Deanery, and in London, appetite for change began earlier, resulting in collaboration between senior clinicians and educationalists with experience of both VTS and research into interactive testing.[3] A substantive change in West Midlands recruitment and selection practice occurred, shifting away from traditional 'CV and interview' format. The resulting recruitment centre – a four (long) station OSCE - was successfully piloted (2009 Appendix 1) and implemented (2010 and 2011). The simulated consultation (called *Communication in the clinical setting*) featuring a trained, professional medical role player, was a core element of the new process, and one ultimately retained in the emerging national model. Similarly, in London the first pilot was undertaken in 2008, using a multi-station approach but without simulated consultation.

The use of simulated patients in dental education has been explored[4-7] although the evidence base is less substantial than in medicine where the methodology has been more extensively applied[8] and reviewed.[9] Simulation affords a number of benefits in assessment situations as it provides a valid context in which to look at an applicant's overall consulting ability. A challenge to simulation is the degree to which it generates consistently measurable outcomes. 'It can be subjective ... unless patients are highly trained and calibrated.'[10]

The National DF1 Recruitment Working Group (a working group brought together by London Deanery to oversee the establishment and operations of national foundation recruitment. The group comprised Deans [or alternates] from the Deaneries who hosted recruitment centres, the Chair of COPDEND [Postgraduate Deans in the UK] and later a British Dental Association Officer, young dentists' representatives and the Chair of the Dental Schools Council) responsible for implementing a national model for DF1 recruitment supported the inclusion of a simulated patient station, recognising the opportunity to assess each applicant holistically. However, unlike the earlier West Midlands scheme – where the role players summatively scored each applicant's interpersonal skills and professionalism – it was decided that the simulated patients in the first national scheme would not formally contribute to scoring the interviews. The rationale was that while psychometric data existed for some simulated patient teams, it did not exist for others. Therefore, only the scores of the observing general dental practitioner (GDPs) 'counted' towards the applicants' results. The simulated patients were nonetheless asked to complete score sheets explicitly for the purposes of establishing the feasibility of their actually awarding scores in future recruitment events. This paper presents the data collected in the first round of national DF1 recruitment centres (2012), with particular focus on how clinical examiners and trained simulated patients rate applicants. The aim, therefore, is improved understanding of the effectiveness and consistency of clinical and simulated patient scoring in foundation level recruitment. It is hoped that outcomes will be of interest to others using, or considering using, simulated patients as scorers in high stakes assessments.

METHOD

The study was an educational evaluation (as such ethical approval was not required. Analysing the consistency of any novel testing process is a routine educational requirement). Data analysed were:

1. The scores actually awarded to each applicant – out of 20 – for performance on the *Communication in the clinical setting* station by each of two independently observing clinicians (GDPs). Clinicians did not confer on the scores
2. The scores recorded for (but not awarded to) each applicant – out of 20 – for performance on the *Communication in the clinical setting* station by the simulated patient.

Guidance and descriptors for different levels of score were provided on the interview marking schedules (Appendix 2). The simulated patient score encompassed five domains – empathy, problem solving, information sharing, communication and professional attitude. For obvious reasons only the clinicians (and not the simulated patients) scored on clinical knowledge and management.

The clinical assessors were experienced dental foundation trainers and programme directors. All had taken part in previous foundation training recruitment.

The same scenario – extraction request from a patient with high blood pressure and a heart murmur – ran at all five national centres over a week, and was up to ten minutes in duration. Simulated patients and clinical assessors were recruited locally at each centre. They were briefed and calibrated at local training events using live and video examples.

Data were retrospectively analysed to test the hypotheses that there would be no significant scoring differences between centres and that inter-rater reliability, by applicant, between paired clinicians, and between clinicians and simulated patients would be strong.

Statistical methods used included correlation coefficients (Pearson and Spearman as appropriate) to examine associations between scores, kappa to examine agreement between outcomes, chi-squared tests of differences in outcomes between centres and one-way analysis of variance to investigate differences between centres. Additionally, Spearman correlations between domain scores and total scores (both clinical and simulated patient) were calculated to assess whether any domains appeared particularly influential.

Geographical centres are anonymised in the text (Centres A-D).

RESULTS

Data were returned by four (of the five) centres. Overall there were 896 applicants represented in this set. Clinician assessor scores for the data collated for this research were missing for 27 applicants. Simulated patient scores were missing for nine. One applicant had a recorded clinician score of 30, which is out of range. This was amended to 20 for analysis purposes.

1. Analysis of the clinician assessor scores

Two independent clinician assessor ratings (out of 20) were available for each applicant. The Pearson correlation between the ratings was 0.752, while the rank correlation was 0.736. There was a high level of agreement between the clinicians' scores.

The difference between clinicians' scores was computed, and extreme values of this

difference (greater than or equal to nine in absolute value) were identified for just six applicants, too small a number for further comparisons.

For the purpose of this pilot study (although not for the overall recruitment process) the average of the two clinician scores was computed and applicants classified as 'passing' if the average score was greater than or equal to ten, 'failing' if the average score was less than or equal to eight, and 'borderline' otherwise. On this classification the six applicants identified as outliers comprised four passes, one fail and one borderline score (see Table 1).

Differences between centres in mean average scores were not statistically significant (one-way ANOVA $F_{3.865} = 2.57$; $p > 0.05$). However, using the classification of scores into pass, fail and borderline, differences between centres were found to be statistically significant ($\chi^2 = 16.68$, df = 3; $p = 0.011$).

Table 2 shows that the proportions of clear passes were higher for Centre B and Centre C than for Centre A and Centre D. This is in line with values in the previous table of means.

2. Analysis of the simulated patient scores

Applicants were assigned a score between 0 and 4 in each of five areas by the simulated patients, so the range of scores was the same as the range of the average of the clinician assessor scores. Simulated patient scores were missing for nine of the applicants (administrative error). Applicants were classified as fail, pass and borderline in the same way as for the clinician score averages.

Table 3 reports the mean simulated patient scores for each of the 4 contributing centres.

One-way analysis of variance shows that these differences are statistically significant ($F = 4.43$, df = 3, 883; $p < 0.005$) but the pattern of differences is not the same as for the clinician scores. In particular Centre A and Centre C have the highest and lowest average simulated patient scores, whereas for the clinicians Centre B and Centre A had the highest and lowest averages respectively.

The results in terms of pass rates based on simulated patient scores – Table 4 – do not show statistically significant differences between centres ($\chi^2 = 4.94$, df = 3; $p > 0.05$).

Table 1 Mean clinician assessor scores by centre

Centre	Mean	N	Std deviation
Centre A	12.8507	201	3.96927
Centre B	13.7021	240	3.82622
Centre C	13.0640	242	3.08574
Centre D	12.9032	186	3.84269
Total	13.1565	869	3.68405

Table 2 Clinician assessor scores – pass, fail and borderline by centre

Fail borderline			Clinical score result			Total
			Pass			
Centre	Centre A	Count	29	13	159	201
		% within centre	14.4%	6.5%	79.1%	100.0%
	Centre B	Count	19	17	204	240
		% within centre	7.9%	7.1%	85.0%	100.0%
	Centre C	Count	13	23	206	242
		% within centre	5.4%	9.5%	85.1%	100.0%
	Centre D	Count	26	18	142	186
		% within centre	14.0%	9.7%	76.3%	100.0%
Total		Count	87	71	711	869
% within centre			10.0%	8.2%	81.8%	100.0%

Table 3 Mean simulated patient scores by centre

Centre	Mean	N	Std deviation
Centre A	14.1765	204	3.79243
Centre B	13.2254	244	3.37176
Centre C	12.9395	248	3.91824
Centre D	13.4607	191	3.78324
Total	13.4149	887	3.73763

Table 4 Simulated patient scores – pass, fail and borderline by centre

Fail borderline			Role player result			Total
			Pass			
Centre	Centre A	Count	17	9	178	204
		% within centre	8.3%	4.4%	87.3%	100.0%
	Centre B	Count	22	9	213	244
		% within centre	9.0%	3.7%	87.3%	100.0%
	Centre C	Count	27	13	208	248
		% within centre	10.9%	5.2%	83.9%	100.0%
	Centre D	Count	26	6	159	191
		% within centre	13.6%	3.1%	83.2%	100.0%
Total		Count	92	37	758	887
% within centre			10.4%	4.2%	85.5%	

3. Comparison of average clinician assessor scores and simulated patient scores

The Pearson correlation coefficient between simulated patient and clinician assessor average scores was 0.61, which while clearly statistically significant ($p < 0.001$) would not be considered as indicating particularly strong agreement. The average difference in scores between simulated patient ratings and clinician ratings was not statistically significant at 0.25 with the simulated patient score higher.

In terms of the classification of applicants into pass, fail and borderline categories the results of the two scoring methods are illustrated in Table 5.

Table 5 shows statistically significant but only moderate agreement (kappa = 0.37, $p < 0.001$).

As the table is a symmetric contingency table the test for differences is based on the symmetrically corresponding off-diagonal cells (that is cells where disagreement is observed). The appropriate test – Bowker's extension of the McNemar test – indicated statistically significant differences. ($\chi^2 = 15.73$, df = 3; $p < 0.01$). Examination of the table shows that simulated patient scoring led to more clear passes and fewer borderline results than clinician scoring.

Overall, 4% of applicants deemed an overall clear pass by clinician ratings were awarded a fail on simulated patient communication ratings. Of those deemed a clear pass on the basis of simulated patient ratings, 5.2% failed on the clinicians' ratings. The table also shows that the most striking difference was in the 'borderline' proportions (8.1% for clinicians *vs.* 4.2% for simulated patients).

Agreement between clinical raters was also examined, and the kappa value was 0.53 (s.e. 0.03, $p < 0.001$) indicating moderate agreement.

Overall the data revealed 147 instances of exact agreement between the two clinician assessors, and 199 cases in which the simulated patient and one clinician agreed exactly. On 25 occasions all three awarded an identical score.

Table 5 Comparison of simulated patient and clinician assessor scores – pass, fail and borderline

Fail borderline			Role player result			Total
			Pass			
Clinical score result	Fail	Count	43	5	38	86
		% within clinical score result	50.0%	5.8%	44.2%	100.0%
		% within role player result	47.8%	13.9%	5.2%	10.0%
	Borderline	Count	19	8	43	70
		% within clinical score result	27.1%	11.4%	61.4%	100.0%
		% within role player result	21.1%	22.2%	5.8%	8.1%
	Pass	Count	28	23	655	706
		% within clinical score result	4.0%	3.3%	92.8%	100.0%
		% within role player result	31.1%	63.9%	89.0%	81.9%
Total		Count	90	36	736	862
% within clinical score result			10.4%	4.2%	85.4%	100.0%	
% within role player result			100.0%	100.0%	100.0%	100.0%	

Table 6 Clinical communication score domains

	Domain average score (s.e.)	Correlation with clinician score	Correlation with simulated patient score
Empathy	2.52 (0.85)	0.46	0.78
Problem solving	2.69 (0.90)	0.52	0.88
Information sharing	2.72 (0.89)	0.53	0.86
Communication (skills)	2.73 (0.81)	0.48	0.85
Professional attitude	2.74 (0.86)	0.53	0.89

4. Domains of the simulated patient score

Table 6 gives the Spearman rank correlations between the clinician assessor average score, the simulated patient total score and each of the individual domains of the simulated patient score. For example, the correlation between the empathy domain score and the clinician average score was 0.46, whereas the correlation between the simulated patient score and the empathy domain score was 0.78. Pearson correlation coefficients were not considered appropriate as the range of the individual scores on the simulated patient elements was restricted. The table shows communication domains, of which only one is 'communication skills' in the sense of items like paraphrasing, questioning style, avoidance of jargon etc.

There was no indication that any particular domain was preponderantly predictive of either the clinical score or the total simulated patient score, although the empathy element had the lowest correlation with both the simulated patient and the clinician scores.

DISCUSSION

Competition demands a robust process, not just here but in most high stakes contexts. The pilot study generated thought-provoking data, and is being used by the national DF1 Recruitment Working Group to implement change. While applicant success was not exclusively on a numeric pass rate (it also included meeting some essential criteria, for example, a 'red flag' alert of unacceptable practice/behaviour might overrule reasonable scores for other areas) for the vast majority of applicants it was the awarded scores that determined outcome. The terms 'pass' and 'fail' are used as shorthand internally for 'success in meeting the recruitment criteria' or not,

but it should be remembered that this is an interactive interview rather than an 'examination'. That is to say assessment in this context means, as with other job interviews, 'assessing the applicant's suitability for the post'.

On this occasion the same clinical scenario was used in each centre as previous internal work generated no evidence of advantage or disadvantage under similar circumstances. The process of the clinical communication interaction was really the subject of interest, rather than the content. However, for future rounds it was decided to use multiple scenarios to create a bank going forward, as this is perceived as good practice.

The high level of agreement between the interviewing clinicians' scores is encouraging. Paired clinicians made very similar judgements about applicants they both observed. This, in conjunction with the level of agreement between simulated patients and clinical assessors, makes a case for going forward with one clinician assessor and one simulated patient actively scoring. This retains the benefits of two independent scorers, but changes the dynamic. Data for this initiative will be harvested from future national recruitments, and reported.

In just five cases (out of 896) a pair of clinicians differed by scores greater than or equal to nine. In these cases, interestingly, the simulated patient score was clearly aligned with one or other of the clinicians. Every assessment has some outliers. This is a very small number, and such examples could be retrospectively moderated at the interviewers' meeting where a grade boundary is affected. There were no cases of differences between a simulated patient score and a clinician score exceeding nine marks, which bodes well for future consistency.

The 147 instances of exact agreement between the two clinician assessors, 199 cases where the simulated patient and one clinician agreed and 25 occasions where the three awarded identical scores bode well. Having a higher number of cases where simulated patient and clinician agreed than where two clinicians agreed suggests that simulated patients have a good shared understanding with their clinical counterparts of what performance merits what score.

There are two caveats to be aware of. Firstly, these results were attained with simulated patient teams with experience of assessing in the medical context, and experience of working in an educational context. Consistent high quality training is paramount for all national assessors, clinical and lay,[11] to establish a shared understanding of performance levels, expected capabilities, recording of objectified evidence, usage of the scoring domains and so forth. For 2012-13 the simulated patients and clinicians will receive comparable training (generated from a single source) to improve standardisation even further.

Secondly, it is important to be mindful that exact agreement between simulated patients and clinical observers is not necessarily the 'gold standard'. Clinical scorers and patient representatives will inevitably be tuned into some different things – a theme widely discussed at national and international educational symposia. As obvious examples the 'patient' can feel the subtleties of the received impact of communication in a way that an external observer often cannot. The clinical observer will be sensitive, in addition to the interpersonal interaction, of clinical safety and be scoring on additional components (eg clinical facts) that the simulated patient will not be judging. While a shared sense of applicant performance is important, in our 22 years of experience the differences can be important too. In the real world the trainee is experienced from the perspective of their trainer and of their patients, and these groups have both shared and different priorities. The inclusion of clinical and lay assessors captures both. Indeed if agreement were identical in all cases we might question the need for two scorers, as the different perspectives add richness to the assessment, and provide moderation.

On that basis the next round of DF1 recruitment will include scores from a single observing clinician assessor and the trained simulated patient (in place of the current two clinical scores). To start the national initiative the two simulated patient teams with the greater consistency in the data set will co-deliver across all six centres signed up for the next round of recruitment. Having a smaller number of assessors trained intensively and assessing often is anecdotally considered good educational practice.

An interesting finding was the tendency of simulated patients to offer more clear scores and fewer borderline scores than clinical staff. Scores in assessments tend to cluster round the median mark. The question of whether a mark of, say, 50/100 is a genuine '50' or an expression of doubt on a fail decision remains unanswered, but confidence in using the full scoring range – including fail grades where performance or patient safety is a concern – should routinely be encouraged in all assessor training.

While the mean average scores actually awarded by centre were not significantly different (Table 1), as expected there was some variance between centres in terms of the score range. This could be by chance, or reflect that national centres can have different demographics, influenced by factors such as the type of applicants attending. Some centres may, for example, include more 'second attempt' applicants, or applicants with more homogenous or diverse prior teaching experiences. One centre, for example, had more non-UK nationals and non-UK graduate applicants. Not planning to capture these data in advance was a limitation, so future differences are being addressed in the ongoing evaluation. As examples examiner online moderation has been set up, and applicant cluster variance is being looked at longitudinally 2012-14.

While mean scores overall did not differ between centres, pass rates sometimes did However, this difference showed in the clinician rather than in the simulated patient score (which had no obvious difference in pass rates by centre). This is suggestive of clinical knowledge being a key influencer on pass rate. It will be interesting to analyse ongoing cumulative data after assessor calibration through training to see if these early, tentative trends are replicated. Awarded scores by centre will also be part of a separate and substantial future analysis. An interesting point for thought is that the objective (knowledge) domains might vary more in score than the more subjective (communication) domains. The latter usually attract more concern in terms of parity.

Analyses unique to the simulated patients were the clinical communication domains. As the simulated patients do not award a clinical score their focus is on the (holistic) communication, so we asked them

to break this down by domain. In terms of the evidence base this exercise generated useful findings. It was to be expected that the domain scores would correlate highly with the total simulated patient scores as the former were included in the latter (note that the domain scores did not involve clinician scoring). Empathy, however, had the weakest correlation with the overall score, while professional attitude had the strongest correlation. Information sharing also was strong. This supports current thinking that observable 'skills' (such as demonstration of empathy) are not the greatest determinant of a successful outcome for the patient. Attitude and other qualities, including expertise, have greater impact (Table 6). The lower correlations by domain with overall clinician score were anticipated, as they had split focus marking both communication and clinical competencies.

Overall these pilot results were interesting and encouraging, supporting the case for simulated patient scores to be counted in future DF1 recruitment schemes. This is not only educationally important (the applicant's full range of skill is addressed, including their patient's need, in line with patient-centred approaches to teaching and learning) but is resource efficient, as the inclusion of a lay scorer reduces the number of clinical staff removed from patient care on a given day. While finance is not the main driver, cost effectiveness may be regarded as a side benefit. If the simulated patient is there anyway they arguably should be used to maximum benefit. Conclusions here will hopefully inform discussion about, and confidence in, using simulated patients as scorers in other dental and medical educational assessment contexts. The aim of establishing internal consistency was met, and lessons learned have already been incorporated into the ongoing scheme, including generation of a larger rotating question base, national standardisation of scorer calibration, and the use of a single simulated patient team (based on strongest data here) to work across all national centres.

1. Eva W, Rosenfield J, Reiter H, Norman G R. An admissions OSCE: the multiple mini interview. *Med Educ* 2004; **38:** 314–326.
2. The National Recruitment Office for GP Training. Online information available at www.gprecruitment.org.uk/ (accessed July 2013).
3. Wiskin C M, Allan T, Skelton J. Hitting the mark: negotiated marking and performance factors in the communication skills element of the VOICEs examination. *Med Educ* 2003; **37:** 22–31.
4. Johnson J A, Kopp K C, Williams R G. Standardized patients for the assessment of dental candidates' clinical skills. *J Dent Educ* 1990; **54:** 331–333.
5. Croft P, White D, Wiskin C, Allan T. Evaluation by dental candidates of a communication skills course using professional simulated patients in a UK school of dentistry. *Eur J Dent Educ* 2005; **9:** 2–9.
6. Gorter R C, Eijkman M A. Communication skills training courses in dental education *Eur J Dent Educ* 1997; **1:** 143–147.
7. Logan H L, Muller P J, Edwards Y, Jakobsen J R. Using standardized patients to assess presentation of a dental treatment plan. *J Dent Educ* 1999; **63:** 729–737.
8. Barrows H S. An overview of the uses of standardized patients for teaching and evaluating clinical skills. *Acad Med* 1993; **68:** 443–451.
9. May W, Part J H, Lee J P. A ten-year review of the literature on the use of standardized patients in teaching and learning: 1996–2005. *Med Teach* 2009; **31:** 487–492.
10. Kramer G A, Albino J E, Andrieu S C *et al.* Dental student assessment toolbox. *J Dent Educ* 2009; **73:** 12–35.
11. Glassman P A, Luck J, O'Gara E M, Peabody J W. Using standardized patients to measure quality: evidence from the literature and a prospective study. *Jt Comm J Qual Improv* 2000; **26:** 644–653.

Appendix 1 Summary results of 2009 pilot (not published)

Three hundred and thirty-five candidates over six days. Little variance was found between paired clinical interviewers scoring the same task. SPs scored using a fuller range of marks than clinicians, with differences in the fail and high end categories. Simulation and structured clinical reasoning tasks were statistically more likely to flag poor performance. Candidate scores overall showed a normal educational distribution. Satisfaction ratings from candidates, on the day, were in excess of 94% agreement for every category of the interactive process (relevance, fairness, information provided and staff performance). Candidate gender as a variable did not reach statistical significance. Data protection rules prohibited analysis of ethnicity.

Appendix 2 Score sheet simulation station (NB clinical assessors captured clinical management safety issues separately)

DF1 Recruitment – CLINICAL COMMMUNICATION SCENARIO

BRIEF GIVEN TO CANDIDATE: See your attached notes. The candidate will have undertaken a role play simulation with you.
Please score based on the following categories. Remember, the 'red card' can be used irrespective of the numeric score awarded.

Empathy and sensitivity. Including respect for your feelings, putting you at ease, clear focus on your situation, and understanding/responding to your worries.	**Comments:**
Reassurance/problem solving about now and the future. Did the dentist's explanation clarify the situation for you? Did s/he solve the problem? Were you involved in (to an *appropriate degree* for you) management plan?	
Information sharing, including the comprehensibility and credibility of what your we told. Also consider clarity of language (avoidance of jargon). Did you understand the management plan?	
Generic clinical communication, including non-verbal messages, appropriate questioning, listening skills, acknowledgement, checking, summary etc.	
Appropriate professional attitudes. Did the dentist inspire trust and confidence? Did s/he demonstrate professionalism? Were you respected?	