JOURNAL OF
Surgical Simulation

Journal of Surgical Simulation (2022) **9,** 127–137
DOI: https://doi.org/10.1102/2051-7726.2022.0017

ORIGINAL ARTICLE

Evaluating the role of simulated practical skills assessments in the UK neurosurgical national selection process

Tom J. Grundy,[a,*] Thomas W. Fallows,[b] Peter C. Whitfield[c] and Ian D. Kamaly-Asl[d]

[a]Manchester Centre for Clinical Neurosciences, Salford Royal Foundation Trust, Stott Lane, Salford, M6 8HD, UK; [b]Manchester, UK; [c]Department of Neurosurgery, University Hospitals Plymouth NHS Trust, Derriford Road, Plymouth, PL6 8DH, UK; [d]Department of Neurosurgery, Royal Manchester Children's Hospital, Oxford Road, Manchester, M18 9LW, UK

**Corresponding author at: Manchester Centre for Clinical Neurosciences, Salford Royal Foundation Trust, Stott Lane, Salford, M6 8HD, UK. Email: tjgpub@gmail.com*

Date accepted for publication: 13 July 2022

Abstract

Background: The selection and training of surgeons is an important undertaking for recruitment and development of a neurosurgical workforce. Meaningful and reliable assessment of practical skills can be problematic due to several factors. We aimed to appraise the performance of candidates in the national neurosurgical recruitment exercise to establish correlation with specific practical skill simulations to assess the value of each simulation in discriminating between candidates. **Methods:** We reviewed 1078 anonymised candidate records between 2009 and 2018 and analysed the prospectively collected Objective Structured Assessment of Technical Skills (OSATS) scores. Statistical evaluation of correlation between simulated assessment scores was performed to establish potential relationships between scoring domains. Correlation coefficients were calculated to establish potential relationships between skill simulation domains. A cohort of candidates was assessed for stereoblindness, and score domains were compared against this trait to assess the role of three-dimensional vision in candidates' performance. **Results:** Significant correlation was noted between specific skill simulation assessments, including microsurgical bead manipulation and suture placement tasks ($\rho = 0.155$, $P < 0.001$); 94% of applicants demonstrated stereoscopic vision. Stereo-blind candidates did not perform significantly poorer in tasks requiring visuospatial skills. A significant degree of correlation was observed between skill domains across assessments suggesting reliability in the testing methods. **Conclusions:** OSATS are a reproducible and established method of assessing simulated surgical skills. Specific scoring domains demonstrated high degrees of correlation, suggesting that they test similar skills. Stereovision does not appear to affect candidates' scores, potentially a result of practical compensation; therefore they may be deemed equally appropriate candidates for neurosurgical selection.

Keywords: *neurosurgery; OSATS; national selection; stereovision; practical skills*

Introduction

The selection and training of surgeons is an important undertaking for the maintenance and development of a surgical workforce that is fit for purpose both in today's neurosurgical practice and in the future. Neurosurgical training is a prolonged and costly process and thus it is vital to select candidates who are suitable, safe and demonstrate aptitude for the specialty.[1,2] Consequently, the aim of any selection process is to identify and select candidates who have greater potential to develop as trainee and consultant practitioners who will be competent and effective neurosurgeons. Historically, selection processes have focused on the applicants' academic achievements, subjective scoring at interview and references submitted.[1]

Academic achievement before assessment is often deemed to be a suitable indicator of general aptitude, scholarly disposition and ability to perform in subsequent surgical training programmes. This paradigm has been shown to be problematic in part due to the omission of practical and technical assessments, despite the obvious fact that technical procedures are an integral part of surgical practice.[2–4]

The development of the United Kingdom national selection programme in 2008 sought to apply an objective, score-based selection model to the recruitment of neurosurgical trainees. Longlisting and subsequent shortlisting processes then allowed for candidates to be examined centrally using a range of assessment techniques. Objective testing undertaken in specific 'stations' such as interview, portfolio review, clinical case management and patient interaction simulation allowed for a validated and credible method to individually assess candidates and compare scores objectively.[2,4]

The introduction of three simulated Objective Structured Assessment of Technical Skills (OSATS) stations into the neurosurgical selection process introduced a practical element to the overall assessment of a candidate's suitability for surgical training.[5,6] The use of OSATS is a well-established, robust and valid measure of practical ability within the limits of the task undertaken.[7,8] Favourable outcomes in simulated OSATS exercises for current surgical trainees have been found to correlate strongly with surgical and procedural ability in practice.[9–13]

Stereoscopic vision and visual discrimination are traits that are believed to be of use to surgeons in performing specific tasks related to dexterity, microsurgery and hand-eye co-ordination.[14] A number of causes for a lack of stereovision are known, the most common of which is amblyopia – otherwise known as a 'lazy eye'. There remains a paucity of evidence to confirm with certainty that a lack of stereoacuity itself is a predictor of poor surgical performance in surgical simulations.[15] The current literature suggests that most surgeons have stereoscopic vision; a notable minority are amblyopic, however they are able to perform surgical procedures to a high standard. The ability to identify potential surgeons with a high degree of stereoacuity, correlated with a high degree of surgical performance in simulated tasks, may be a useful aspect of surgical selection.[16]

Assessment of procedural and technical ability at time of assessment and recruitment to surgical specialty training is a key aspect in determining a candidate's suitability for neurosurgical training. Prolonged and complex assessments can be problematic for candidates, assessors and the entire system of selection, however the process is required to be sufficiently detailed to select the appropriate candidates. Our study is one of the first to evaluate the role of simulated surgical tasks in the recruitment of neurosurgical trainees.

We analysed the simulated practical skills assessment stations within the ST1 recruitment process to investigate the type of simulation or assessment that would allow maximal meaningful discrimination between ability of candidates. We also aimed to assess the degree of stereovision held by neurosurgical applicants to measure the potential association with surgical task performance.

Methods

Ethical considerations

Before the assessment process, written permission was obtained from each candidate for the use of anonymised data for quality control and research purposes.

The stations

The three OSATS stations featured in the neurosurgery ST1 recruitment were initially designed to simulate and test specific surgical skills likely to be encountered by junior trainees in a neurosurgical run-through training programme, some of which would likely be familiar to candidates and others that the candidate may not have encountered before. Candidates were scored against a global assessment rating matrix. The OSAT stations were derived and adapted for neurosurgical tasks from previously validated objective scoring matrixes described and validated by Faulkner et al.[13] The stations are described below.

Microscope simulation

The candidate was provided with a Zeiss S7 microscope, a range of surgical instruments, a bead tray with a pattern and an array of coloured beads. Candidates were allowed a fixed period of time to set up the microscope for use, select the appropriate instruments and subsequently move the coloured beads into the bead tray in keeping with the coloured pattern provided. Candidates were scored objectively in the following domains: precision and accuracy, use of time and motion, knowledge of instruments, hand-eye co-ordination and use of the microscope. The total number of beads correctly placed was also counted.

Brainlab simulation biopsy

The candidate was provided with a simulation head with a pre-designed craniotomy in the right frontal area. The Brainlab setup, registration and trajectory had been pre-set for each candidate. Candidates were given a demonstration of how to use the instrument to target the biopsy needle to the biopsy site. Candidates were given time to practice the procedure with guidance from the invigilator. Candidates were then asked to recreate the demonstration under timed conditions. Candidates were scored in the following domains: precision, use of time and motion, appropriate use of the instruments, hand-eye co-ordination and flow of the procedure. Time taken to reach the target in seconds was also recorded.

Suturing simulation

The candidate was provided with a high-fidelity mannequin featuring a pre-made skin incision of set length, a range of surgical instruments and a selection of sutures. The candidate was instructed to achieve surgical closure of the wound.

Candidates were scored in the following domains: respect of tissues, use of time and motion, handling of instruments, placement of sutures and securing of knots.

Scoring and standardisation of assessment

Station assessors were trained in task performance and OSAT scoring system processes before the assessment sessions. Independent supervision and benchmarking was undertaken for all new assessors. Assessment sessions were randomly selected for quality assurance supervision and there was minimal interchange of assessment personnel in each assessment year. The structured scoring schemes using for OSAT assessment are presented in the Appendix 1.

Stereoscopic vision

The assessment of stereoacuity was not a formal test station that contributed to a candidate's score. Information was gathered with the candidate's consent for research purposes. The candidates were formatively tested using the TNO standardised assessments (Laméris Ootech BV, Ede, Netherlands) to identify the presence and discrimination of stereoscopic vision.

Anonymised candidate score sheets were reviewed for all candidates invited to the neurosurgical application interviews between 2008 and 2019. Candidate scores were collated by score within each scoring domain for the three stations. Comparisons were made between candidates' scores between domains and stations.

Statistical analysis

Combined domain percentage scores for each station were calculated (total number of beads and total time were not included). Combined station scores for each candidate were subsequently plotted against each other to display the general correlation between station scores.

Individual scores were ranked using the average rank function in Microsoft Excel. From the ranked data, Spearman's rank correlation coefficient (ρ) was calculated using the following formula: $\rho = 1 - (6\sum d^2 n^3 - n)$, where d is the difference in rank between two stations for a given candidate and n is the number of candidates. Two-tailed P values were calculated using the following formula: $\mathrm{t} - \mathrm{stat} = \rho\sqrt{n} - 21 - \rho^2$. A heatmap was generated using Microsoft Excel conditional formatting to visually demonstrate statistical correlation; green demonstrates greater correlation and red demonstrates poorer correlation.

Average scores from each practical skills station were plotted against stereoscopic visual trait and a Mann-Whitney U test was applied to calculate statistical significance. To assess the impact of previous surgical training on score test outcomes, suture scores from ST1 candidates were compared with those of ST3 candidates using Student's paired t test.

Results

We reviewed 1078 anonymised candidate examination records from between 2009 and 2018. The data collection methods varied in 371 records. This was related to an alteration in the scoring domains for one station over the 10-year period. Records with common scoring domains that were not directly comparable with the scores for subsequent years were excluded from the analysis to ensure that appropriate comparisons and correlations could be calculated from the scoring domains across the remaining dataset. The examination records demonstrated a normal distribution of mean scores in practical skills assessments.

Demographic data were not available for reporting due to the anonymised nature of the data collection.

Station specific scores

The median microscope score was 20 of a possible 25 points. The lowest recorded score was 5 and the maximum score was 25. The mean bead count was 47.65. There was a significant correlation found between the average score and the mean bead count ($\rho = 0.58$, $P < 0.01$).

The median Brainlab score was 18 of a possible 25 points. The lowest recorded score was 5 and the maximum score was 25. The mean time taken to complete the outlined task was 45.24 s. Similarly, statistically significant negative correlation was noted between the Brainlab score and the time taken to perform the test ($\rho = -0.69$, $P < 0.01$). Negative correlation between these parameters indicates that candidates who gained higher scores in this station also completed the task in a shorter time.

The mean suture score was noted to be 20 of a possible 30 points. It was assumed that suturing would be the most practiced of the skills tested using the stations outlined. A sampled comparison of suture scores was undertaken to assess potential differences between the scores of ST1 and ST3 applicants. There was a significant difference in the performance of ST1 applicants compared with their more experienced colleagues. The median ST1 score was 20 and the mean of the ST3 sample was 24 ($P < 0.01$) using a two-tailed Student's t test.

Correlation between stations

Assessment of correlation between tested domains revealed that there was positive correlation between the overall microscope score and the overall suture score (Fig. 1).

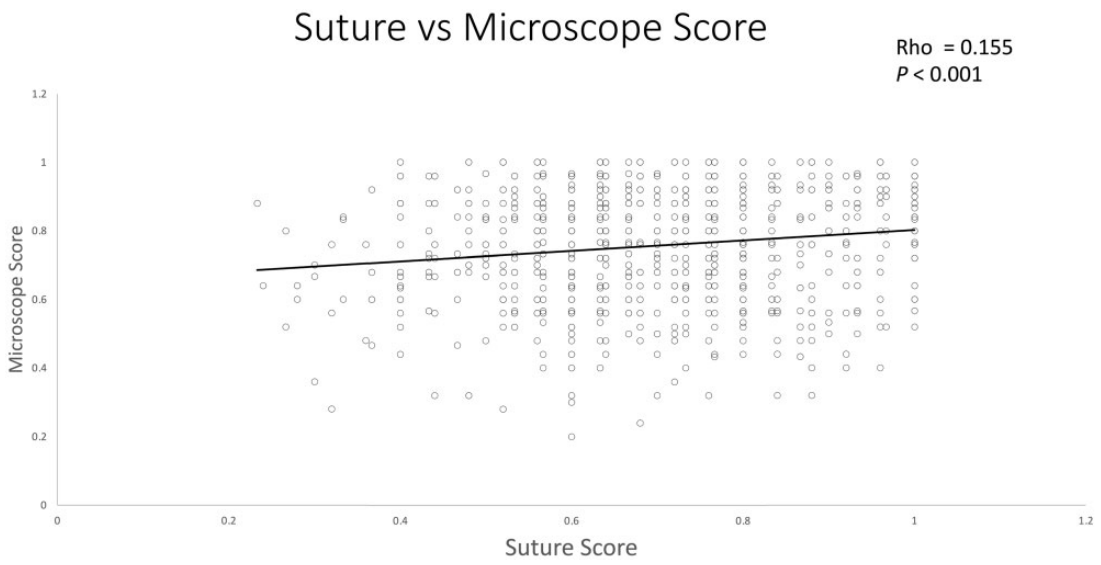


Figure 1. Scatter plot demonstrating relative suture score versus relative microscope bead score. Spearman's rho correlation coefficient ρ = 0.155.

The correlation of specific tasks and stations was assessed using the ranking correlation described previously. As expected, domain scores within specific stations demonstrated high degrees of correlation. The Brainlab score was compared with the Brainlab time ($\rho = -0.685$, $P < 0.001$). Similarly a higher score within the microscope station resulted in a strong likelihood of a high bead score ($\rho = 0.577$, $P < 0.001$; Fig. 2).

A favourable overall score in the microscope station correlated with a favourable overall score in the suturing station ($\rho = 0.155$, $P < 0.001$). This was similarly reflected when the microscope bead score was compared with the suture score ($\rho = 0.155$, $P < 0.001$). Brainlab station scores did not correlate significantly with any of the measures noted from the other practical skill stations. The correlation coefficients between compared scores are shown in more detail in Fig. 3.

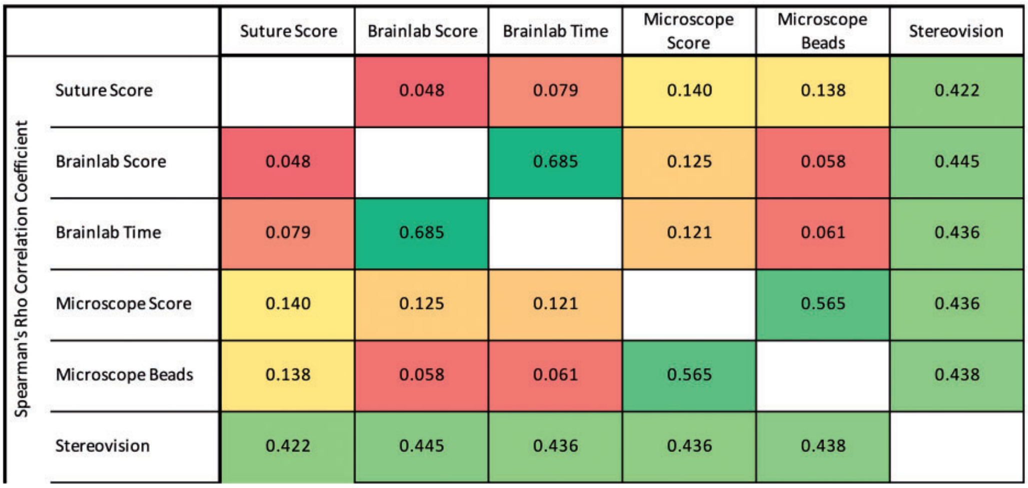
On formal assessment, 665 (94.1%) of applicants who were tested were found to have stereoscopic vision. The presence of stereoscopic vision was presumed to be advantageous in the performance of practical or visuo-spatial tasks. The presence of stereoscopic vision did show a correlation with the mean Brainlab score; however, on statistical testing using a Mann-Whitney U test, it failed to reach statistical significance. The average Brainlab score for stereoscopic candidates was 0.73 compared with 0.67 for amblyopic applicants (Fig. 4).

Discussion

This is a large study of prospectively recorded surgical simulation OSATS scores from 10 years of established practice in the assessment of neurosurgical applicants. Review of candidate performance within and between practical skills assessment stations demonstrated important correlations between specific scoring domains that could inform the future structure of candidate assessments.

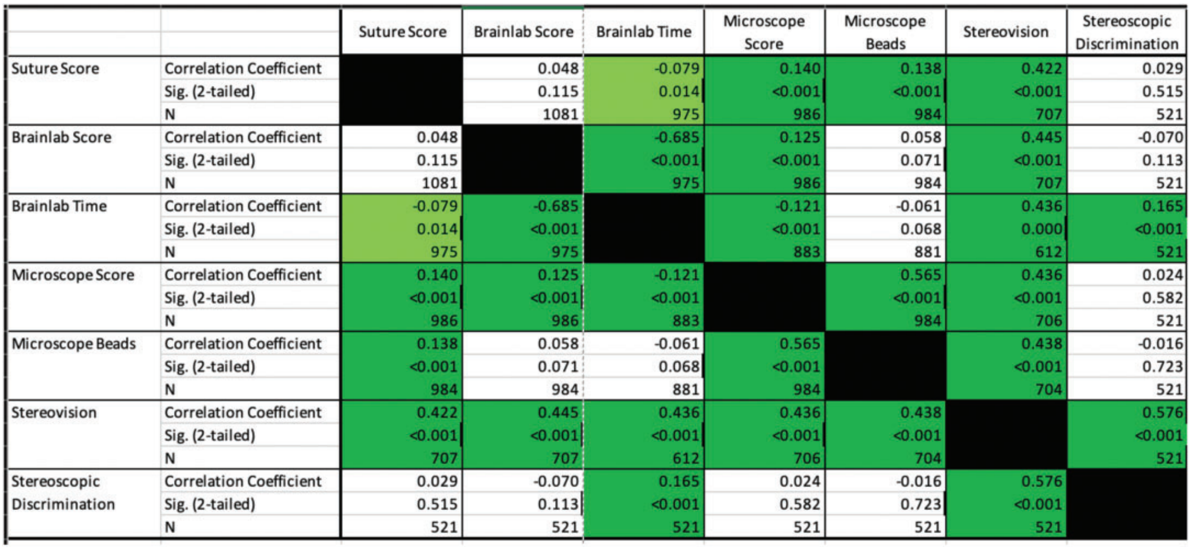
The practical skills assessment stations were designed pragmatically to test skills that were believed to be relevant and important to neurosurgical training. Generally, scores within each station domain demonstrated a high degree of correlation with other markers of success in the station. The total score in the microscope station had a strong correlation with the number of beads the candidate was able to move during the assessment. This correlation was expected on the assumption that these domains were essentially testing the same skill, namely dexterity when using a microscope. Our study demonstrated good reliability and correlation of scores within stations in comparison with other similar published studies.[17,18] Likewise, the scoring domains were comparable with respect to the generalisability and specific validity for the tasks outlined.[19,20]

Suturing skills have been considered to be a learned experience and thus it was expected that candidates with more experience in suturing would score higher in this station. To test this hypothesis, we assumed that applicants for ST3 roles would have a higher degree of experience and therefore competence in suturing ability. This finding is reproduced across the literature. Oliver[18] demonstrated that practiced surgical tasks perform better in OSATS assessment. Similarly Panait et al.[21] found that among general surgical trainees, those with 2–3 years more surgical experience showed notably higher OSATS scores in surgical



		Suture Score	Brainlab Score	Brainlab Time	Microscope Score	Microscope Beads	Stereovision
Spearman's Rho Correlation Coefficient	Suture Score		0.048	0.079	0.140	0.138	0.422
	Brainlab Score	0.048		0.685	0.125	0.058	0.445
	Brainlab Time	0.079	0.685		0.121	0.061	0.436
	Microscope Score	0.140	0.125	0.121		0.565	0.436
	Microscope Beads	0.138	0.058	0.061	0.565		0.438
	Stereovision	0.422	0.445	0.436	0.436	0.438	

Figure 2. Heatmap demonstrating relative Spearman's rho correlation coefficients of average simulated station scores and stereovisual trait.



		Suture Score	Brainlab Score	Brainlab Time	Microscope Score	Microscope Beads	Stereovision	Stereoscopic Discrimination
Suture Score	Correlation Coefficient		0.048	-0.079	0.140	0.138	0.422	0.029
	Sig. (2-tailed)		0.115	0.014	<0.001	<0.001	<0.001	0.515
	N		1081	975	986	984	707	521
Brainlab Score	Correlation Coefficient	0.048		-0.685	0.125	0.058	0.445	-0.070
	Sig. (2-tailed)	0.115		<0.001	<0.001	0.071	<0.001	0.113
	N	1081		975	986	984	707	521
Brainlab Time	Correlation Coefficient	-0.079	-0.685		-0.121	-0.061	0.436	0.165
	Sig. (2-tailed)	0.014	<0.001		<0.001	0.068	0.000	<0.001
	N	975	975		883	881	612	521
Microscope Score	Correlation Coefficient	0.140	0.125	-0.121		0.565	0.436	0.024
	Sig. (2-tailed)	<0.001	<0.001	<0.001		<0.001	<0.001	0.582
	N	986	986	883		984	706	521
Microscope Beads	Correlation Coefficient	0.138	0.058	-0.061	0.565		0.438	-0.016
	Sig. (2-tailed)	<0.001	0.071	0.068	<0.001		<0.001	0.723
	N	984	984	881	984		704	521
Stereovision	Correlation Coefficient	0.422	0.445	0.436	0.436	0.438		0.576
	Sig. (2-tailed)	<0.001	<0.001	<0.001	<0.001	<0.001		<0.001
	N	707	707	612	706	704		521
Stereoscopic Discrimination	Correlation Coefficient	0.029	-0.070	0.165	0.024	-0.016	0.576	
	Sig. (2-tailed)	0.515	0.113	<0.001	0.582	0.723	<0.001	
	N	521	521	521	521	521	521	

Figure 3. Correlation matrix demonstrating correlative indices between stations, including total numbers and *P* values.

instrument tasks compared with surgical applicants or even internal medical trainees. Other similarly designed studies have also replicated these trends.[22] The statistically significant difference in scores between these ST1 and ST3 applicants highlights a potential inequality in assessment, however the separate application tracks for these two roles precludes any direct competition between these groups.

The correlation of scores between stations may represent testing of similar skills across different stations. For example, a high score in the microscope station was a strong predictor for a favourable score in the suture station. Moreover, a high microscope bead count was the strongest predictor of success in suturing ability. This is potentially explained by both tasks requiring strong performance in manual dexterity; however, this has not yet been reproducibly established in the wider literature.[22] Despite the strong performance of OSATS in the assessment of practical skills assessment,[7,23,24] the use of an independent single score such as the microscope bead count may be an appropriate surrogate for OSATS assessment in this particular assessment station.

Performance of candidates in the microscope station did not demonstrate any meaningful correlation with scores in the Brainlab station. The assumption that these two stations are a test of different skill sets may be a possible explanation for this. Undertaking the simulated biopsy requires looking at a screen relaying two-dimensional information and translating that into movements of the candidates' hands in three-

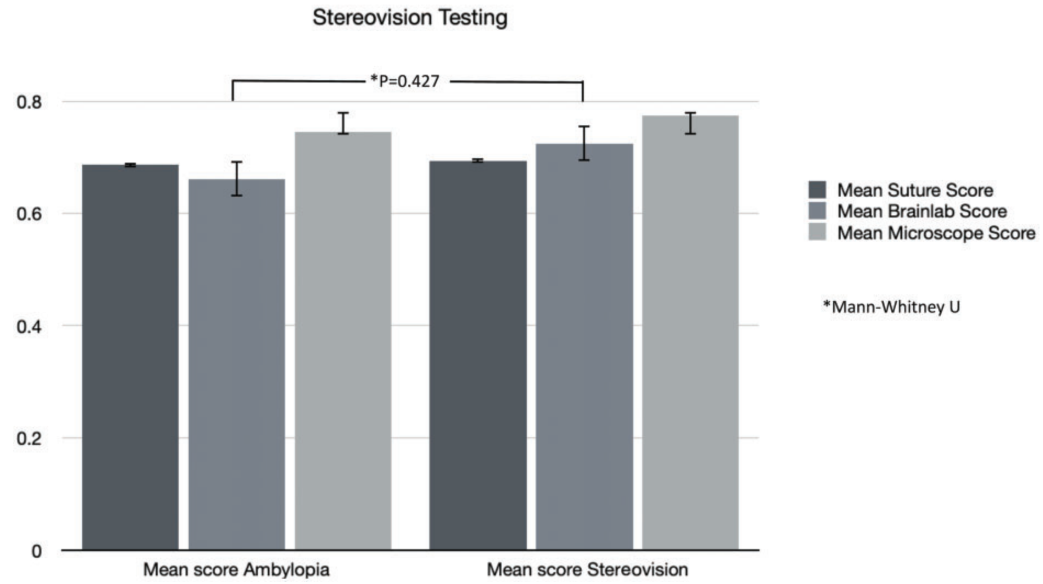


Figure 4. Bar chart demonstrating the mean suture simulation score in the presence of stereovisual trait. Mann-Whitney U test of significance calculated.

dimensional space. This is different to the microscope and suture skills where hand movements are based on direct visual feedback but clearly is also a relevant skill in neurosurgery. Testing these different skills within the practical stations is likely to be beneficial in discerning the best candidates.

The trait of stereoblindness, the lack of stereovisual ability, was assumed to be a disadvantageous trait in the performance of practical or surgical skills.[25,26] The study by Nibourgh et al.[15] of medical students performing fine visuo-motor tasks showed a significant difference between groups based on their stereovisual status. Although our data did show a trend towards higher scores in the Brainlab station for those with stereovision, this failed to reach statistical significance. Given that amblyopia is a relatively rare trait in the population, it may be that a greater sample size is required to demonstrate a true effect.[27] However, a more intriguing possibility is that no statistical disadvantage was found because surgical applicants with a stereoscopic deficiency may have an innate disadvantage but have made use of compensatory mechanisms to perform similarly to applicants with stereovision. A study by Fergo et al.[16] noted a stereoblindness rate of nearly 10% in established surgeons, suggesting this trait may not significantly impair surgical ability.

This study makes use of a large dataset collected over a 10-year period. This provides a large number of data points to consider, but a number of small refinements have been made to the scoring and data collection process over this period. This introduces potential disparity in the datasets from year to year. To mitigate the effects of this disparity on our study, those scoring domains that were not common to all year groups have been excluded to preserve the homogeneity of the data.

Limitations of this study include the anonymised nature of the data relating simulated tasks to the wider aspects of assessment including interview and communication skills. In addition, in this study, we were not able to correlate candidate's performance in simulated practical skills with ongoing or future microsurgical ability or performance to measure the 'trainability' of surgical candidates. The main barrier was the anonymised nature of the candidate data, rendering identification and assessments of successful candidates problematic.

Another limitation of this study is the measure of stereoacuity as a physical trait rather that a skill or learned behaviour, therefore caution must be exercised in its use to assess suitability for microsurgical learning.

As part of an ongoing process of quality control for the recruitment process, psychometric analysis reports were commissioned to evaluate the overall correlation of scores and inter-examiner variability. Wider assessment of general interview skills, simulated patient consultations and clinical management interviews demonstrated consistent correlation between non-practical skills stations. Wider assessment of general interview skills, simulated patient consultations and clinical management interviews demonstrated consistent correlation between non-practical skills stations (M Kerrin, R Shaw, confidential report, Neurosurgery National

Selection: 2014 Psychometric Evaluation of Neurosurgical National Selection Centre; M Kerrin, A Aitkenhead, A Smith, confidential report, Neurosurgery National Selection: 2013 Evaluation). In a post hoc survey of candidates, the practical skills stations were thought to be relevant to the practice of neurosurgery, appropriate for the relevant entry level and allowed them sufficient opportunity to demonstrate their practical skills.

This study aimed to review the role of practical simulated skills stations in the recruitment of neurosurgical trainees. Due to the anonymised nature of the dataset, consideration of these practical scores in the wider candidate scoring process has not been investigated in this body of research and remains an area of ongoing research for future publication.

Prospective evaluation of correlation between non-practical skills assessment and performance in practical simulated skills stations is an area of future focus of our research to establish a wider assessment of practical skills assessments in the overall context of applications to the neurosurgical speciality.

Conclusions

This large-scale study of practical skills assessment in neurosurgery recruitment has demonstrated a number of interesting areas for consideration. Over a number of years, simulated practical skills assessments within the neurosurgical national selection process have been shown to be an established, fair and acceptable method of appraising candidates' practical abilities with respect to surgical skill. This has been found to be acceptable to both candidates and assessors.

The clear difference in scores between ST1 and ST3 applicants within the suturing station demonstrates a likely point of disparity and inequality between candidates applying between these two levels. It may be prudent to design a modified or different surgical task for ST3 applicants to perform to adjust for previous surgical experience and to discriminate between candidates' abilities.

Correlation between specific station scores, for example, bead count and suture score, may be explained at least in part by testing of the same or similar skills. Future work to further assess the similarities between these stations may allow future assessment frameworks to be streamlined to yield maximum information on more limited assessment.

The link between stereovisual ability and performance of practical tasks involving depth perception is not fully understood. Although this study demonstrates a general correlation between stereoblindness and reduced scoring, it failed to reach significance. A further review including more specific testing and a greater number of participants will form part of future work.

This study is an example of the advancing role of simulation in the area of surgical speciality recruitment and selection. This process has been successfully deployed over a prolonged time period and has demonstrated consistent value in selection of neurosurgical candidates. We aim to build on this work to compare candidates' practical scores with overall candidate assessment in neurosurgical recruitment to give further insights into the wider role of surgical skills assessment in recruitment.

Conflict of interest

The authors have no conflicts of interest to declare.

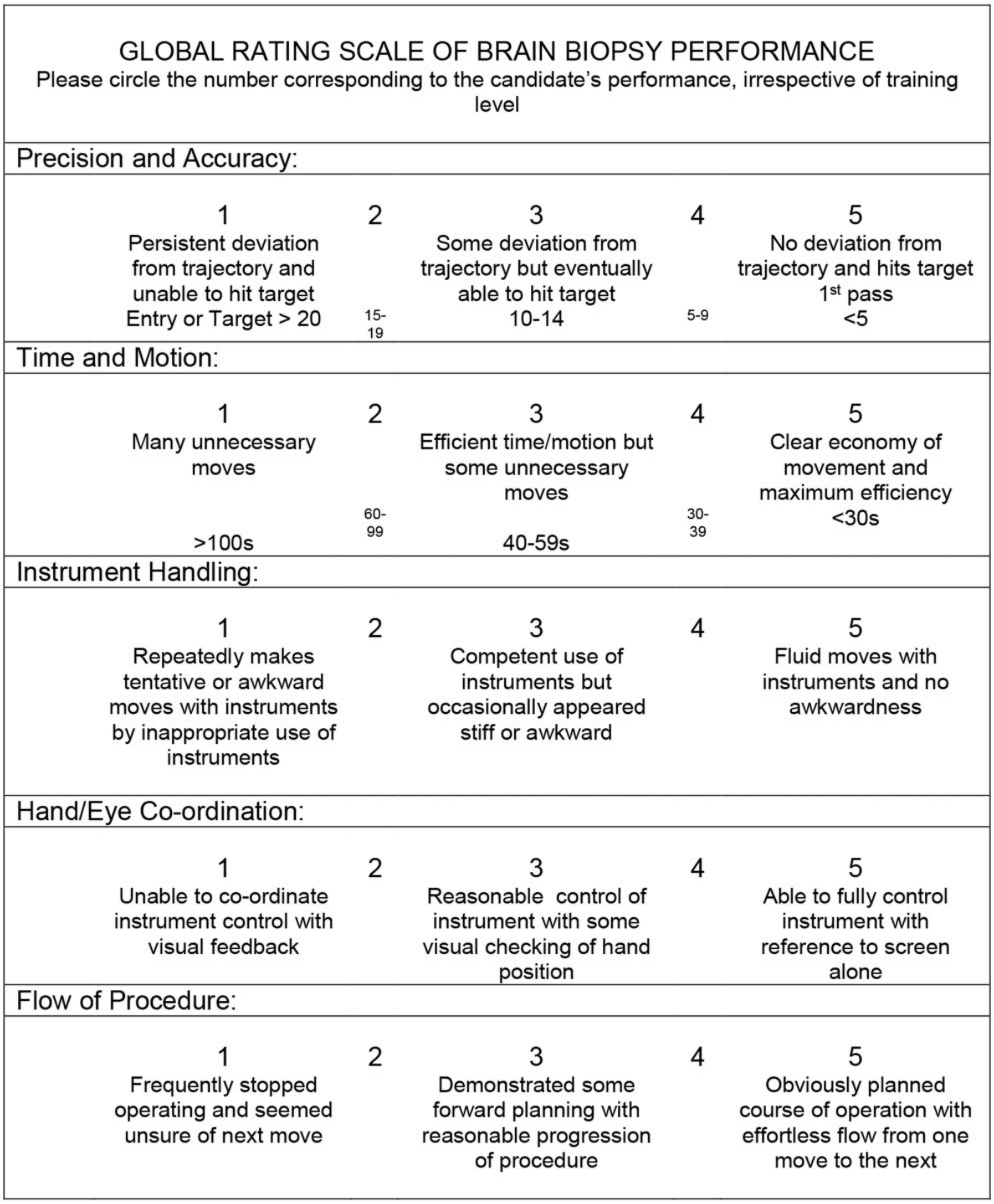
References

1. Carroll SM, Kennedy AM, Traynor O, Gallagher AG. Objective assessment of surgical performance and its impact on a national selection programme of candidates for higher surgical training in plastic surgery. J Plastic Reconstr Aesthetic Surg 2009; 62(12): 1543–1549. https://doi.org/10.1016/j.bjps.2008.06.054.
2. Alamri A, Chari A, McKenna G, Kamaly-Asl I, Whitfield PC. The evolution of British neurosurgical selection and training over the past decade. Med Teach 2018; 40(6): 1–5. https://doi.org/10.1080/0142159X.2018.1444744.
3. Bann S, Darzi A. Selection of individuals for training in surgery. Am J Surg 2005; 190(1): 98–102. https://doi.org/10.1016/j.amjsurg.2005.04.002.
4. Zhang L, Kamaly I, Luthra P, Whitfield P. Simulation in neurosurgical training: a blueprint and national approach to implementation for initial years trainees. Brit J Neurosurg 2016; 30(5): 577–581. https://doi.org/10.1080/02688697.2016.1211252.
5. Bisson DL, Hyde JP, Mears JE. Assessing practical skills in obstetrics and gynaecology: educational issues and practical implications. Obstetrician Gynaecol 2006; 8(2): 107–112. https://doi.org/10.1576/toag.8.2.107.27230.
6. Beard JD. Assessment of Surgical Skills of Trainees in the UK. Ann Royal Coll Surg Engl 2008; 90(4): 282–285. https://doi.org/10.1308/003588408X286017.
7. Martin JA, Regehr G, Reznick R, Macrae H, Murnaghan J, Hutchison C, et al. Objective structured assessment of technical skill (OSATS) for surgical residents. Br J Surg 1997; 84(2): 273–278. https://doi.org/10.1046/j.1365-2168.1997.02502.x.
8. Niitsu H, Hirabayashi N, Yoshimitsu M, Mimura T, Taomoto J, Sugiyama Y, et al. Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale

to evaluate the skills of surgical trainees in the operating room. Surg Today 2013; 43(3): 271–275. https://doi.org/10.1007/s00595-012-0313-7.

9. Schuwirth LW. Assessing medical competence: finding the right answers. Clin Teach 2004; 1(1): 14–18. https://doi.org/10.1111/j.1743-498X.2004.00012.x.
10. Bould MD, Crabtree NA, Naik VN. Assessment of procedural skills in anaesthesia. Br J Anaesth 2009; 103(4): 472–483. https://doi.org/10.1093/bja/aep241.
11. Datta V, Bann S, Aggarwal R, Mandalia M, Hance J, Darzi A. Technical skills examination for general surgical trainees. Br J Surg 2006; 93(9): 1139–1146. https://doi.org/10.1002/bjs.5330.
12. Sultana CJ. The Objective Structured Assessment of Technical Skills and the ACGME Competencies. Obstet Gyn Clin N Am 2006; 33(2): 259–265. https://doi.org/10.1016/j.ogc.2006.01.004.
13. Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. Acad Med 1996; 71(12): 1363–1365. https://doi.org/10.1097/00001888-199612000-00023.
14. Schoob A, Kundrat D, Kahrs LA, Ortmaier T. Stereo vision-based tracking of soft tissue motion with application to online ablation control in laser microsurgery. Med Image Anal 2017; 40: 80–95. https://doi.org/10.1016/j.media.2017.06.004.
15. Nibourg LM, Wanders W, Cornelissen FW, Koopmans SA. Influence of stereoscopic vision on task performance with an operating microscope. J Cataract Refract Surg 2015; 41(9): 1919–1925. https://doi.org/10.1016/j.jcrs.2014.12.066.
16. Fergo C, Burcharth J, Pommergaard H-C, Rosenberg J. Age is highly associated with stereo blindness among surgeons: a cross-sectional study. Surg Endosc 2016; 30(11): 4889–4894. https://doi.org/10.1007/s00464-016-4826-9.
17. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004; 38(9): 1006–1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x.
18. Oliver C. Developing and maintaining an assessment system - a PostGraduate Medical Education Training Board (PMETB) guide to good practice. London: Postgraduate Medical Education and Training Board; 2007.
19. Lammers RL, Davenport M, Korley F, Griswold-Theodorson S, Fitch MT, Narang AT, et al. Teaching and assessing procedural skills using simulation: metrics and methodology. Acad Emerg Med 2008; 15(11): 1079–1087. https://doi.org/10.1111/j.1553-2712.2008.00233.x.
20. Lefor AK, Harada K, Dosis A, Mitsuishi M. Motion analysis of the JHU-ISI Gesture and Skill Assessment Working Set II: learning curve analysis. Int J Comput Ass Rad 2021; 16(4): 589–595. https://doi.org/10.1007/s11548-021-02339-8.
21. Panait L, Larios JM, Brenes RA, Fancher TT, Ajemian MS, Dudrick SJ, et al. Surgical skills assessment of applicants to general surgery residency. J Surg Res 2011; 170(2): 189–194. https://doi.org/10.1016/j.jss.2011.04.006.
22. Rethans J, Norcini JJ, Barón-Maldonado M, Blackmore D, Jolly BC, LaDuca T, et al. The relationship between competence and performance: implications for assessing practice performance. Med Educ 2002; 36(10): 901–909. https://doi.org/10.1046/j.1365-2923.2002.01316.x.
23. Carroll SM, Kennedy AM, Traynor O, Gallagher AG. Objective assessment of surgical performance and its impact on a national selection programme of candidates for higher surgical training in plastic surgery. J Plastic Reconstr Aesthetic Surg 2009; 62(12): 1543–1549. https://doi.org/10.1016/j.bjps.2008.06.054.
24. Michelson JD, Manning L. Competency assessment in simulation-based procedural education. Am J Surg 2008; 196(4): 609–615. https://doi.org/10.1016/j.amjsurg.2007.09.050.
25. Cagenello R, Halpern DL, Arditi A. Binocular enhancement of visual acuity. J Opt Soc Am 1993; 10(8): 1841. https://doi.org/10.1364/JOSAA.10.001841.
26. Goodwin RT, Romano PE. Stereoacuity degradation by experimental and real monocular and binocular amblyopia. Invest Ophthalmol Vis Sci 1985; 26(7): 917–923.
27. Chopin A, Bavelier D, Levi DM. The prevalence and diagnosis of ‘stereoblindness’ in adults less than 60 years of age: a best evidence synthesis. Ophthalmic Physiol Opt 2019; 39(2): 66–85. https://doi.org/10.1111/opo.12607.

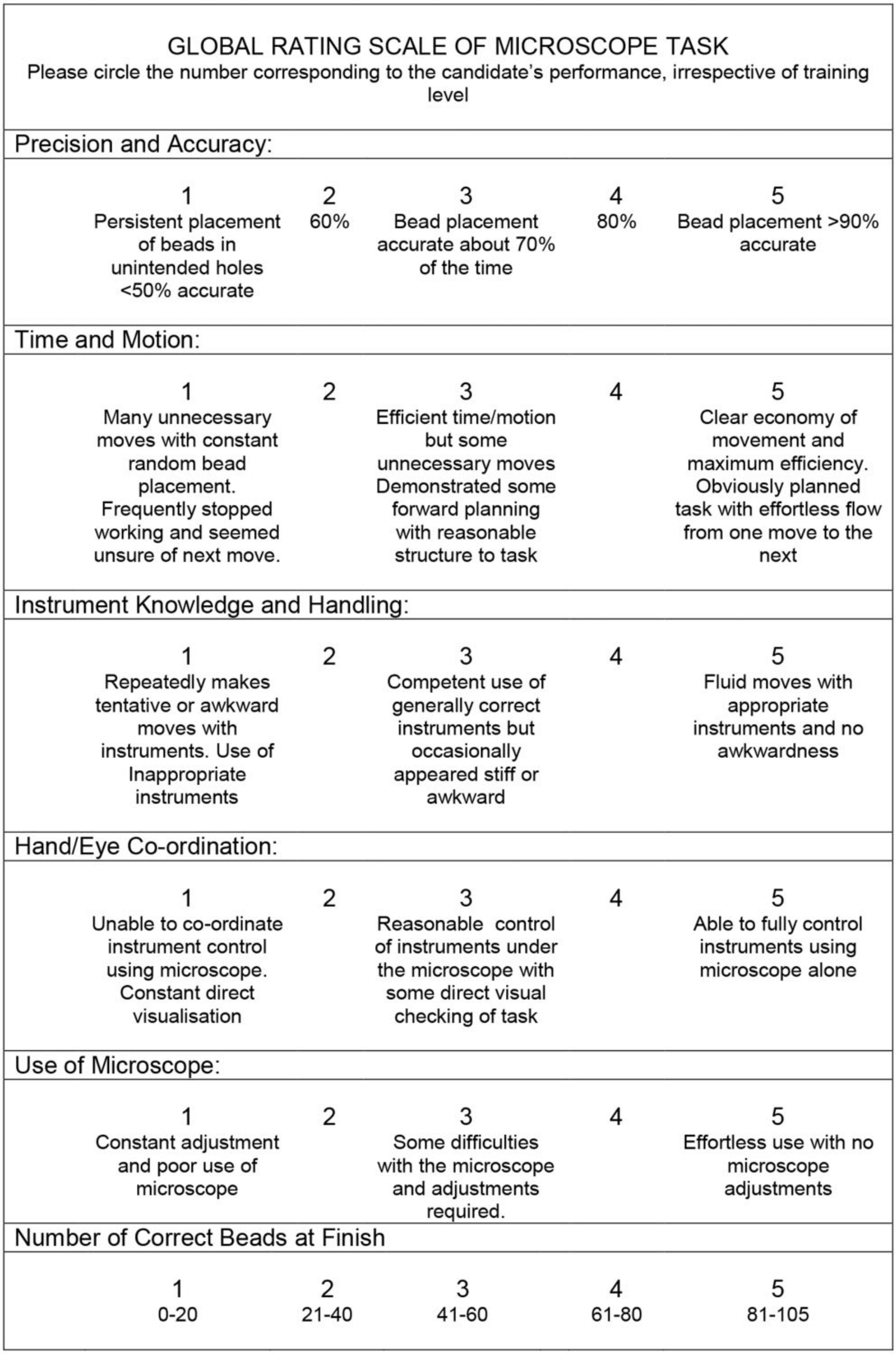
Appendix 1



GLOBAL RATING SCALE OF BRAIN BIOPSY PERFORMANCE

Please circle the number corresponding to the candidate's performance, irrespective of training level

Precision and Accuracy:				
1	2	3	4	5
Persistent deviation from trajectory and unable to hit target		Some deviation from trajectory but eventually able to hit target		No deviation from trajectory and hits target 1st pass
Entry or Target > 20	15-19	10-14	5-9	<5
Time and Motion:				
1	2	3	4	5
Many unnecessary moves		Efficient time/motion but some unnecessary moves		Clear economy of movement and maximum efficiency
>100s	60-99	40-59s	30-39	<30s
Instrument Handling:				
1	2	3	4	5
Repeatedly makes tentative or awkward moves with instruments by inappropriate use of instruments		Competent use of instruments but occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
Hand/Eye Co-ordination:				
1	2	3	4	5
Unable to co-ordinate instrument control with visual feedback		Reasonable control of instrument with some visual checking of hand position		Able to fully control instrument with reference to screen alone
Flow of Procedure:				
1	2	3	4	5
Frequently stopped operating and seemed unsure of next move		Demonstrated some forward planning with reasonable progression of procedure		Obviously planned course of operation with effortless flow from one move to the next



GLOBAL RATING SCALE OF MICROSCOPE TASK

Please circle the number corresponding to the candidate's performance, irrespective of training level

Precision and Accuracy:

1	2	3	4	5
Persistent placement of beads in unintended holes <50% accurate	60%	Bead placement accurate about 70% of the time	80%	Bead placement >90% accurate

Time and Motion:

1	2	3	4	5
Many unnecessary moves with constant random bead placement. Frequently stopped working and seemed unsure of next move.		Efficient time/motion but some unnecessary moves Demonstrated some forward planning with reasonable structure to task		Clear economy of movement and maximum efficiency. Obviously planned task with effortless flow from one move to the next

Instrument Knowledge and Handling:

1	2	3	4	5
Repeatedly makes tentative or awkward moves with instruments. Use of Inappropriate instruments		Competent use of generally correct instruments but occasionally appeared stiff or awkward		Fluid moves with appropriate instruments and no awkwardness

Hand/Eye Co-ordination:

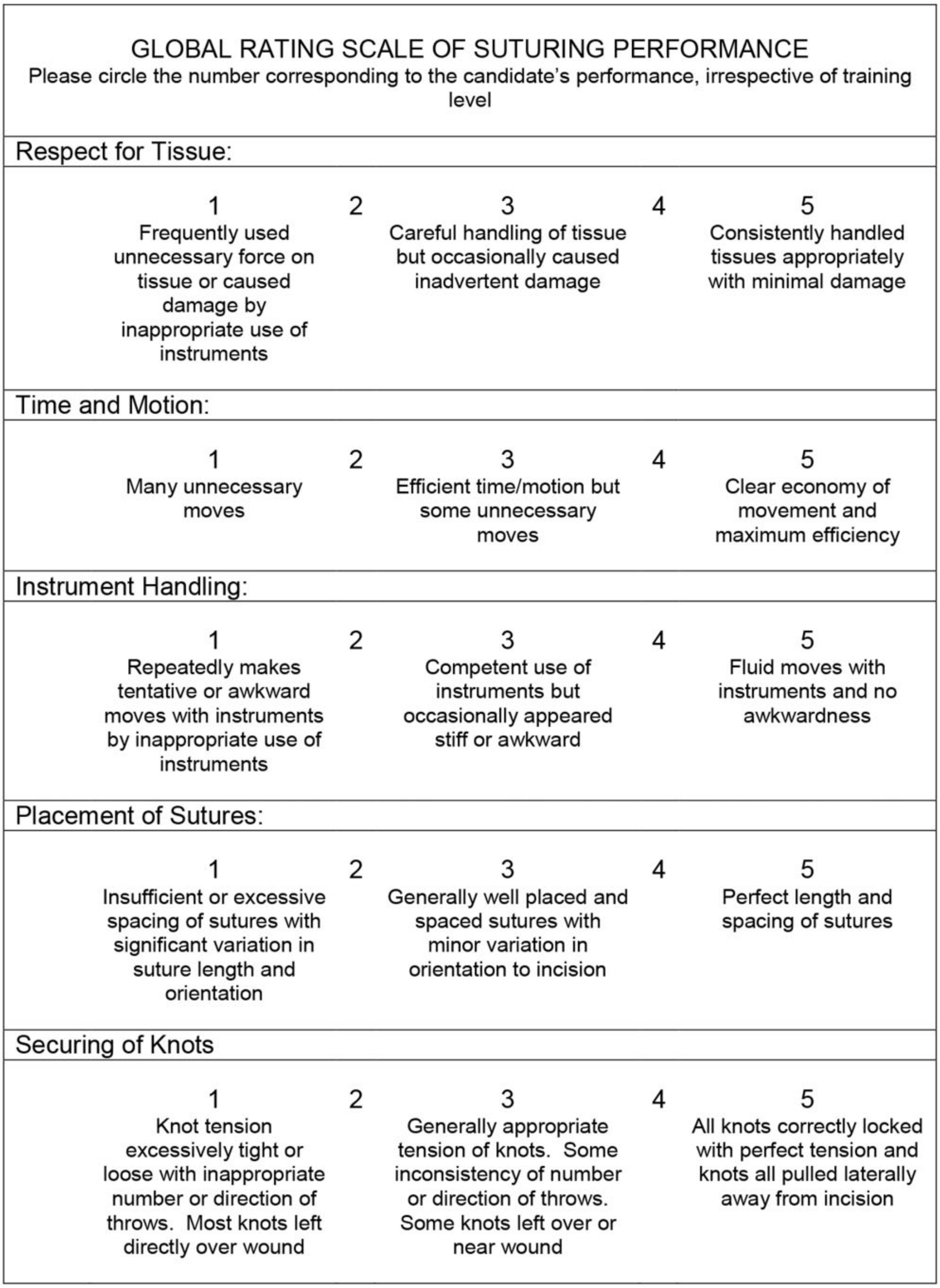
1	2	3	4	5
Unable to co-ordinate instrument control using microscope. Constant direct visualisation		Reasonable control of instruments under the microscope with some direct visual checking of task		Able to fully control instruments using microscope alone

Use of Microscope:

1	2	3	4	5
Constant adjustment and poor use of microscope		Some difficulties with the microscope and adjustments required.		Effortless use with no microscope adjustments

Number of Correct Beads at Finish

1	2	3	4	5
0-20	21-40	41-60	61-80	81-105



GLOBAL RATING SCALE OF SUTURING PERFORMANCE

Please circle the number corresponding to the candidate's performance, irrespective of training level

Respect for Tissue:				
1	2	3	4	5
Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments		Careful handling of tissue but occasionally caused inadvertent damage		Consistently handled tissues appropriately with minimal damage
Time and Motion:				
1	2	3	4	5
Many unnecessary moves		Efficient time/motion but some unnecessary moves		Clear economy of movement and maximum efficiency
Instrument Handling:				
1	2	3	4	5
Repeatedly makes tentative or awkward moves with instruments by inappropriate use of instruments		Competent use of instruments but occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness
Placement of Sutures:				
1	2	3	4	5
Insufficient or excessive spacing of sutures with significant variation in suture length and orientation		Generally well placed and spaced sutures with minor variation in orientation to incision		Perfect length and spacing of sutures
Securing of Knots				
1	2	3	4	5
Knot tension excessively tight or loose with inappropriate number or direction of throws. Most knots left directly over wound		Generally appropriate tension of knots. Some inconsistency of number or direction of throws. Some knots left over or near wound		All knots correctly locked with perfect tension and knots all pulled laterally away from incision