A Comparison of Visual Rating Scales and Simulated Virtual Reality Metrics in Neurosurgical Training: A Generalizability Theory Study

Alexander Winkler-Schwartz[1], Ibrahim Marwa[1], Khalid Bajunaid[1,3], Muhammad Mullah[2], Fahad E. Alotaibi[1,4], Abdulgadir Bugdadi[1,5], Robin Sawaya[1], Abdulrahman J. Sabbagh[3,6], Rolando Del Maestro[1]

■ BACKGROUND: **Adequate assessment and feedback remains a cornerstone of psychomotor skills acquisition, particularly within neurosurgery where the consequence of adverse operative events is significant. However, a critical appraisal of the reliability of visual rating scales in neurosurgery is lacking. Therefore, we sought to design a study to compare visual rating scales with simulated metrics in a neurosurgical virtual reality task.**

■ METHODS: **Neurosurgical faculty rated anonymized participant video recordings of the removal of simulated brain tumors using a visual rating scale made up of seven composite elements. Scale reliability was evaluated using generalizability theory, and scale subcomponents were compared with simulated metrics using Pearson correlation analysis.**

■ RESULTS: **Four staff neurosurgeons evaluated 16 medical student neurosurgery applicants. Overall scale reliability and internal consistency were 0.73 and 0.90, respectively. Reliability of 0.71 was achieved with two raters. Individual participants, raters, and scale items accounted for 27%, 11%, and 0.6% of the data variability. The hemostasis scale component related to the greatest number of simulated metrics, whereas respect for no-go zones and tissue was correlated with none. Metrics relating to instrument force and patient safety (brain volume removed and blood loss) were captured by the fewest number of rating scale components.**

■ CONCLUSIONS: **To our knowledge, this is the first study comparing participant's ratings with simulated performance. Given rating scales capture less well instrument force, quantity of brain volume removed, and blood loss, we suggest adopting a hybrid educational approach using visual rating scales in an operative environment, supplemented by simulated sessions to uncover potentially problematic surgical technique.**

INTRODUCTION

As residency programs continue to evolve toward a competency-based curriculum, there is an increasing need for assessment of resident technical skills. Adequate assessment and feedback remain a cornerstone of psychomotor skills acquisition, particularly within neurosurgery where the consequence of adverse operative events is great.[1] Visual rating scales remain convenient tools for generating organized formative assessments. Different rating scales for surgery have been developed, including the Objective Structured Assessment of Technical Skills (OSATS), which has been used previously in a neurosurgical context.[2-4] A theoretical limitation of visual rating scales is the risk of rater subjectivity in skills assessment. Furthermore, little information exists on the ability of rating scales to capture subtler aspects of performance, including instrument force applied during a procedure. This last point is particularly important because consistent evidence from the neurosurgical

Key words
- Assessment
- Education
- Neurosurgery
- Resident
- Simulation
- Surgery

Abbreviations and Acronyms
OSATS: Objective Structured Assessment of Technical Skills

From the [1]Neurosurgical Simulation and Artificial Intelligence Learning Centre, Department of Neurology & Neurosurgery and [2]Department of Epidemiology, Biostatistics & Occupational Health, McGill University, Montreal, Quebec, Canada; [3]Division of Neurosurgery, Department of Surgery, Faculty of Medicine, University of Jeddah, Jeddah, Saudi Arabia; [4]Neurosurgical Department, National Neuroscience Institute, King Fahad Medical City, Riyadh, Saudi Arabia; [5]Department of Surgery, Faculty of Medicine, Umm Al Qura University, Makkah, Saudi Arabia; and [6]Clinical Skills and Simulation Center, King Abdulaziz University, Jeddah, Saudi Arabia

*To whom correspondence should be addressed: Alexander Winkler-Schwartz, M.D.
[E-mail: manuscriptinquiry@gmail.com]*

*Citation: World Neurosurg. (2019) 127:e230-e235.
https://doi.org/10.1016/j.wneu.2019.03.059*

Journal homepage: www.journals.elsevier.com/world-neurosurgery

Available online: www.sciencedirect.com

simulation literature suggests that applied force differentiates levels of expertise.[5-10] In addition, a recent study found that excess force applied during live neurosurgical operations is associated with increased intraoperative bleeding.[11]

The objective of the project was to conduct a generalizability study to better understand the use of a visual rating scale of operative performance in neurosurgery and to compare it with computerized metrics generated during a virtual reality neurosurgical operative procedure. We hypothesize that both methods will measure the same underlying construct, namely, surgical performance.

MATERIALS AND METHODS

Subjects

Medical student applicants to a single Canadian neurosurgery program in 2015 were recruited to participate in a trial involving a simulated brain tumor resection task.[8] Sixteen of the 17 applicants participated, comprising over 70% of the national neurosurgical applicant pool for that study year.[11] Data were collected at a single time point within the Neurosurgical Simulation and Artificial Intelligence Learning Centre in a controlled laboratory environment void of distracting noise. No follow-up data were collected. All students signed an approved university ethics board consent form before trial participation. All procedures followed were in accordance with the ethical standards of the responsible committee on human experimentation (institutional and national) and with the Helsinki Declaration of 1975, as revised in 2008.

High Fidelity Simulator and Brain Tumor Resection Task

Participant performance during an established virtual reality brain tumor resection task[5] was assessed using construct-validated metrics[6,12] for the NeuroVR (CAE Healthcare, Montreal, Quebec, Canada) simulation platform providing real-time visual and haptic feedback. The results of this analysis are available in a previous publication.[8] Participants were instructed to remove sequentially 6 spherical tumors of identical stiffness and glioma-like color while minimizing damage to simulated normal tissue. Tumor stiffness (Young modulus = 9 kPa) was higher than that of the surrounding normal tissue (Young modulus = 3 kPa) to facilitate the ability of participants to differentiate the tumor—normal tissue interface. The task was completed with an ultrasonic aspirator and suction device held in the dominant and nondominant hand, respectively. See Figure 1 for example.

Performance Video Recording and Rating Scale

Graphical representation of the virtual surgical environment is delivered via computer monitor via the NeuroVR graphic card port. Each eye is presented with an offset view of the operative field, therefore recreating the stereoscopy of a neurosurgical microscope. A high-resolution recording of the virtual reality operation from the perspective of the user was obtained by directing the graphical output in parallel to a DVD recording device.

To reduce potential bias, anonymized participant video recordings were shared with four neurosurgical faculty from two institutions and rated using a modified OSATS Global Rating Scale.[13] The scale is made up of seven composite elements (respect for tissue, economy of movement, instrument handling, overall flow, hemostasis, respect for normal brain, and overall score) measured on a 10-point Likert scale. The scale was produced by the authors after collection of the simulated performance data. Neurosurgical faculty serving as evaluators were not privy to the scale components prior to its use as an evaluation tool in the study.

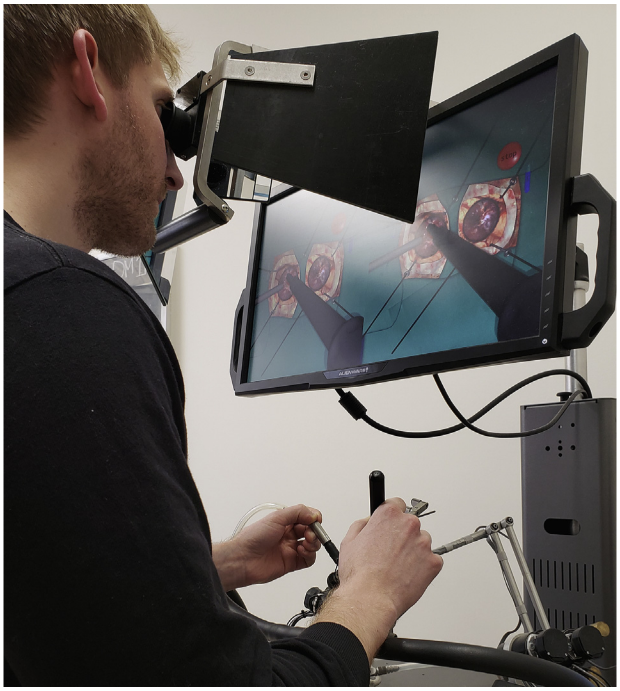

Figure 1. One of the authors performs a simulated brain tumor resection task on the NeuroVR neurosurgical simulation platform.

Statistical Analysis

We report descriptive statistics as counts and percentages for categorical variables. For continuous variables, means and standard deviations are used. Continuous variables include visual rating scale items (respect for tissues, economy of movement, instrument handling, flow, hemostasis, no-go zones, and overall score) and demographic information (number of neurosurgery elective weeks undertaken and number of surgical skin closures performed). Categorical variables include demographic information (previous exposure to simulators). Generalizability theory was used to evaluate scale reliability. G_String with urGENOVA (McMaster Education Research, Innovation & Theory Faculty of Health Sciences, McMaster University, Hamilton, Ontario, Canada) was used to generate variance components of participants, raters, and scale items, and their interactions. Because all raters evaluated every participant, the study design was considered fully crossed.

Simulated metrics from a previous publication[8] were compared with the rating scale subcomponents using Pearson correlation coefficient analysis. For ease of analysis and to further compare the rating scale with simulated metrics, single composite scores for both were created. A rating scale total score was generated

Table 1. Descriptive Statistics of Rating Scale Subcomponents

	Rating	Range	
Scale Item	Mean ± Standard Deviation	Minimum	Maximum
Respect for tissues	4.28 ± 1.80	1	8
Economy of movement	4.12 ± 1.82	1	8
Instrument handling	3.90 ± 1.71	1	8
Flow	4.10 ± 1.78	2	9
Hemostasis	4.30 ± 2.22	1	9
No-go zones	4.75 ± 2.10	1	9
Overall	3.81 ± 1.59	1	8

Four staff neurosurgeons used the rating scale in 64 observations in 16 medical student applicants to neurosurgery residency at McGill University.

by adding individual subcomponents together (range, 7—70). The performance metrics of efficiency index, bimanual forces ratio, suction coordination index, and path length index were combined to create a total metric score (range 0—8). Scores of 0, 1 and 2 were assigned to a given performance metric if an individual achieved below the 25th percentile, between the 25th and 50th percentile, and above the 50th percentile, respectively, compared to their peers. These four metrics were selected because these have shown to best differentiate performance between groups[6] and within groups.[8] Missing data, if present, were replaced with means. All statistical analyses were completed using STATA version 13.0 (StataCorp., LLC, College Station, Texas, USA).

RESULTS

Four staff neurosurgeons evaluated 16 medical students for a total of 64 observations. Table 1 includes a descriptive analysis, demonstrating use of the full range of the Likert scale. Demographic information is available in a previous publication[8] and can be summarized as follows: 7 out of 16 participants (43%) previously used a simulator, the mean number of neurosurgery elective weeks was 11.2 ± 4.6 (range, 4—22), and the mean number of surgical skin closures was 10.9 ± 6.3 (range, 1—25).

Five observations across 3 participants were missing and were replaced with means. Additionally, one reviewer failed to complete the overall scale subcomponent for all participants, representing 16 missing observations. As a result, the overall scale subcomponent was excluded from the generation of the composite rating scale score.

Generalizability theory analysis demonstrated relative g coefficients corresponding to an overall reliability of 0.73 and internal consistency of 0.90. A decision study was conducted, demonstrating that scale reliability of 0.71 can be achieved with only 2 raters (relative error coefficient, and keeping item facet fixed). A single rater failed to complete the overall scale subcomponent of the visual rating scale; therefore, their scores were replaced with those for the group mean.

Table 2. Sources of Variance in Scores

Effect	*df*	T Score	Sum of Squares	Mean Squares	Variance Component	Variance (%)
Participants	15	530.80	530.81	35.39	0.98	27.0
Raters	3	156.21	156.21	52.07	0.39	10.9
Items	6	36.453	36.45	6.08	0.02	0.6
Interactions						
Participants × raters	45	962.96	275.95	6.13	0.75	20.6
Participants × items	90	817.47	250.21	2.78	0.47	13.0
Raters × items	18	241.94	49.280	2.74	0.11	3.2
Participants × raters × items	270	1541.06	242.15	0.90	0.90	24.7

Table 2 displays the variance components associated with the score. Greatest and least sources of data variance were explained by individual participants and individual rating scale items, respectively.

Table 3 represents comparison of the visual rating scale subcomponents with known individual simulated metrics using Pearson correlation analysis. The low variance in the scale items (0.6%) and significant interitem correlation among the scale subcomponents justified the creation of a summative total score for the scale. The scale components with no significant relation to any metrics were respect for no-go zones and respect for tissue (however it should be noted that, even though not significant, they are both negatively correlated with brain volume removed). The scale component which is significantly correlated with the greatest number of simulated metrics is hemostasis, in which positive correlation is seen for efficiency index, suction coordination index, path length index, tumor percentage removed, brain volume removed, sum of forces in the dominant hand, and maximum force dominant hand, and a significant negative correlation with blood loss. The other scale subcomponents have statistically significant correlation with efficiency index, coordination index, path length index, and sum of forces in dominant hand. The only two visual rating scale components that have a significant negative correlation with the bimanual force ratio are instrument handling and overall score. Those metrics relating to instrument force (sum of forces in nondominant hand, maximum force in dominant hand, and maximum force in nondominant hand) and patient safety (brain volume removed and blood loss) were captured by the fewest number of scale subcomponents.

Finally, composite total of visual rating scale score and composite total simulated metric score demonstrated a significant positive correlation (Pearson correlation, 0.31; P = 0.01) (Figure 2). The mean total simulated metric score was 4 ± 2.1.

DISCUSSION

Based on studies of technical performance in neurosurgery, we have recently introduced a conceptual framework to understand

Table 3. Comparison of Scale Subcomponents with Known Simulated Metrics

	Bimanual Cognitive				Quality	Safety					
								Instrument Force			
								Dominant		Nondominant	
	Efficiency Index	Path Length Index	Suction Coordination Index	Bimanual Forces Ratio	Tumor Percentage Removed	Brain Volume Removed	Blood Loss	Sum of Forces	Maximum Force	Sum of Forces	Maximum Force
Hemostasis	0.53	0.41	0.47		0.58	0.39	−0.51	0.45	0.35		
Overall	0.61	0.43	0.46	−0.29	0.44			0.52			
Instrument	0.41	0.33	0.33	−0.37				0.31			−0.29
Economy	0.44	0.36	0.39		0.29			0.39			
Flow	0.65	0.50	0.61		0.47			0.52			
Respect											
No-go											

Pearson correlation coefficients when $P < 0.05$ are shown.

surgical expertise in neurosurgery.[14] Although it is clear that many nontechnical factors, such as clinical decision-making, contribute to expertise, having a framework allows one to better structure research questions relating to the interaction of cognitive and motor domains and how these contribute to operative outcomes, particularly at a challenging juncture in the surgery. In keeping with this, this study aims to further clarify how one may adequately assess technical skills in neurosurgery and to better establish the role for, and limitations of, visual rating scales.

There are a number of strengths related to the visual rating scale. The scale demonstrated overall reliability for as few as two raters.

The main effect for participant's variance component accounts for the largest percentage (27%) of total variability, allowing for generalization of the findings to future potential participants. In assessment, it is desirable for a given scale to capture a large component of variability from participants.[15] Interestingly, these findings recapitulate variability observed in a previous study in the same population, whereby participant performance segregated into three discrete groups: high, middle, and low performers.[8]

The percentage of total subsection variability for participants by items interaction effect (13%) shows that the ratings of participants differ somewhat across scale items, averaged over raters, suggesting perhaps that each item of the scale measures a different aspect of performance. Another interpretation of these findings is that scale performance within an individual is not uniform (i.e., candidates may differ in their relative strengths and weaknesses). In our previous publication, we introduced the concept of technical abilities customized training in neurosurgery, whereby a custom psychomotor intervention tailored to the individual needs of a particular learner is carried out.[8] Given this, it may be interesting to repeat the current study with neurosurgical residents and faculty to evaluate whether performance, as judged by the visual rating scale, becomes more uniform with increasing experience.

The scale item's main effect variance component has a rate of 0.6% in total variance, indicating that participant's ratings on subcomponents of the scale were similar. A low (3.2%) variance component for rater by item interaction effect shows that individual item scales were scored similarly by a given rater.

Weaknesses associated with the scale include the high variance (20.6%) component for participant by rater interaction, suggesting that a given rater scores a given candidate more leniently or severely than other raters. This difference, however, may be accounted for by the fact that the four raters came from four different neurosurgical subspecialties (spine, oncology, epilepsy, and trauma), in addition to perhaps differing comfort with rating participants through video recordings of a simulated performance. Further rater training and calibration may also reduce variability.[16] Additionally, the participant, item, and rater interaction plus further unmeasured sources of variation were high, indicating that up to roughly one quarter of the variability is not explained by the factors measured in the study.

Other observations include a moderate variance component for the rater main effect (10.9%), suggesting that some raters are more lenient than others in their scoring across all candidates (i.e., hawks vs. doves).

Although not the first case of an OSATs inspired checklist's use in neurosurgery,[2-4] this study provides interesting insights on the strengths and limitations of visual rating scales. Contrary to our initial hypothesis, aspects of the visual rating scale specifically included to capture adverse events (avoidance of no-go zone and respect for tissue) were not associated with damage to healthy simulated brain. Furthermore, the visual rating scale was not able to properly determine force characteristics exerted by the

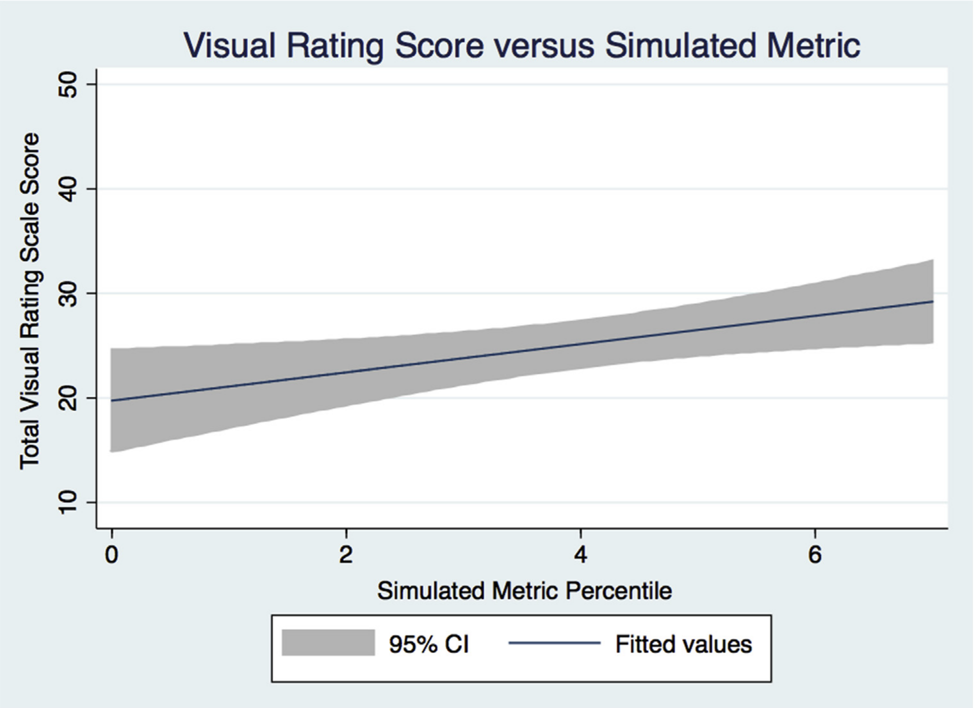

Figure 2. Comparison of composite total visual rating scale score versus composite simulated metric score (Pearson correlation = 0.31, $P = 0.01$). Composite total of scale obtained by summing scale subcomponents (range, 6—60). Composite total of simulated metrics corresponded to below 25th percentile, between 25th and 50th percentile, and above 50th percentile performance on efficiency index, bimanual forces ratio, suction coordination index, and path length index (range, 0—8). Note, overall score subcomponent not included in composite score. CI, confidence interval.

participants. This may be because of the 2-dimensional nature of the video recordings. Although arduous, this limitation could be overcome by providing the evaluators a means of viewing the surgical video in stereoscopy. Similar to the participants using the NeuroVR, evaluators of live surgery obtain a stereoscopic image of the surgical field through the operative microscope, and as such may be better suited to judge deformations in tissue caused by force exerted by a trainee. Interestingly, the positive correlation in the composite scores between the visual rating scale and simulated metrics suggests that both methods may broadly be measuring the same underlying construct, namely technical surgical performance.

The NeuroVR platform allows measurement of force application by individual instruments in all simulated tumor areas, therefore providing a comprehensive 3-dimensional representation of force application during simulated tumor operations. Our group has exploited this information to develop the pyramid and surgical fingerprint concepts, which have contributed to our understanding of the detrimental influence of force application in specific tumor regions.[17,18] These results would suggest that force may be a crucial element to closely monitor during neurosurgical operations. In a recent study by Sugiyama et al.,[11] using force profiles measured by specialized bipolar instruments during neurosurgical operations was associated with increased odds of intraoperative bleeding.

As such, this rating scale may be used to evaluate performance in an operative setting. However, as previously mentioned, if instructors and trainees would like to better understand force applied during a surgical procedure, simulation technology should be used as an adjunct.

These findings come at an important time as resident training is not only being seen as a responsibility of accreditation bodies throughout the world but is increasingly coming under the guise of quality improvement.[19] Simply put, better methods of assessment and training can help reduce patient harm.

Limitations

There are several limitations to this study. First, having raters from various neurosurgical subspecialties may have contributed to differing ratings of individuals. It may not always be feasible to have raters from the same subspecialty available to rate participants. Therefore, this represents a real-world application of this rating scale. Future improvements may lie in selecting a homogeneous rater population more familiar with the evaluated procedure. Second, this study only includes medical students; however, our previous work with simulation suggests that medical students and junior residents share many similar psychomotor characteristics.[5] Third, by virtue of the study design, a performance during a real operation was not rated; however, this has been previously demonstrated by others to be feasible.[3] Finally, by design, no causal relationship can be inferred between the rating scale and simulated metrics; however, the appropriate correlation, as observed for example between the hemostasis subcomponent and blood loss on the simulator, suggests that a similar underlying construct may be evaluated by both systems.

CONCLUSIONS

The visual rating scale can reliably be administered by as few as two raters and seems to reflect operative performance as measured on the simulator. However, force exerted during the neurosurgical operation and the quantity of brain volume removed and blood loss were less well captured by the visual rating scale. To our knowledge, this is the first study to be able to concurrently compare participant's ratings with their computationally measured performance and operative complications. We suggest adopting a hybrid educational approach using visual rating scales in an operative environment, supplemented by simulated training sessions to uncover potentially problematic surgical technique.

ACKNOWLEDGMENTS

The authors thank Dr. Valérie Dory and Dr. Meredith Young for their input. The authors also thank all the medical students and raters who participated in this study and Robert DiRaddo, Group Leader, Simulation, Life Sciences Division, National Research Council of Canada at Boucherville, and his team, including Denis Laroche, Valérie Pazos, Nusrat Choudhury, and Linda Pecora, for their support in the development of the scenarios used in these studies. The authors also thank all the members of the Simulation, Life Sciences Division, National Research Council of Canada.

REFERENCES

1. Jensen RL, Alzhrani G, Kestle JRW, Brockmeyer DL, Lamb SM, Couldwell WT. Neurosurgeon as educator: a review of principles of adult education and assessment applied to neurosurgery. *J Neurosurg*. 2017;127:949-957.

2. Aldave G, Hansen D, Briceno V, Luerssen TG, Jea A. Assessing residents' operative skills for external ventricular drain placement and shunt surgery in pediatric neurosurgery. *J Neurosurg Pediatr*. 2017;19:377-383.

3. Hadley C, Lam SK, Briceño V, Luerssen TG, Jea A. Use of a formal assessment instrument for evaluation of resident operative skills in pediatric neurosurgery. *J Neurosurg Pediatr*. 2015;16:497-504.

4. Sarkiss CA, Philemond S, Lee J, et al. Neurosurgical skills assessment: measuring technical proficiency in neurosurgery residents through intraoperative video evaluations. *World Neurosurg*. 2016;89:1-8.

5. Bajunaid K, Mullah MA, Winkler-Schwartz A, et al. Impact of acute stress on psychomotor bimanual performance during a simulated tumor resection task. *J Neurosurg*. 2016;126:71-80.

6. Alotaibi FE, AlZhrani GA, Mullah MAS, et al. Assessing bimanual performance in brain tumor resection with NeuroTouch, a virtual reality simulator. *Oper Neurosurg*. 2015;11:89-98.

7. Gelinas-Phaneuf N, Choudhury N, Al-Habib AR, et al. Assessing performance in brain tumor resection using a novel virtual reality simulator. *Int J Comput Assist Radiol Surg*. 2014;9:1-9.

8. Winkler-Schwartz A, Bajunaid K, Mullah MA, et al. Bimanual psychomotor performance in neurosurgical resident applicants assessed using NeuroTouch, a virtual reality simulator. *J Surg Educ*. 2016;73:942-953.

9. Alotaibi FE, AlZhrani GA, Sabbagh AJ, Azarnoush H, Winkler-Schwartz A, Del Maestro RF. Neurosurgical assessment of metrics including judgment and dexterity using the virtual reality simulator NeuroTouch (NAJD Metrics). *Surg Innov*. 2015;22:636-642.

10. AlZhrani G, Alotaibi F, Azarnoush H, et al. Proficiency performance benchmarks for removal of simulated brain tumors using a virtual reality simulator NeuroTouch. *J Surg Educ*. 2015;72:685-696.

11. Sugiyama T, Lama S, Gan L, Maddahi Y, Zareinia K, Sutherland GR. Forces of tool-tissue interaction to assess surgical skill level. *JAMA Surg*. 2018;153:234-242.

12. Azarnoush H, Alzhrani G, Winkler-Schwartz A, et al. Neurosurgical virtual reality simulation metrics to assess psychomotor skills during brain tumor resection. *Int J Comput Assist Radiol Surg*. 2015;10:603-618.

13. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Darzi A. Toward feasible, valid, and reliable video-based assessments of technical surgical skills in the operating room. *Ann Surg*. 2008;247:372-379.

14. Sawaya R, Alsideiri G, Bugdadi A, et al. Development of a performance model for virtual reality tumor resections. *J Neurosurg*. 2018;1:1-9.

15. Streiner DL, Norman GR, Cairney J. *Health Measurement Scales: A Practical Guide to Their Development and Use*. Oxford, England, UK: Oxford University Press; 2015.

16. Feldman M, Lazzara EH, Vanderbilt AA, DiazGranados D. Rater training to support high-stakes simulation-based assessments. *J Contin Educ Health Prof*. 2012;32:279-286.

17. Sawaya R, Bugdadi A, Azarnoush H, et al. Virtual reality tumor resection: the force pyramid approach. *Oper Neurosurg*. 2017;14:686-696.

18. Azarnoush H, Siar S, Sawaya R, et al. The force pyramid: a spatial analysis of force application during virtual reality brain tumor resection. *J Neurosurg*. 2017;127:171-181.

19. Pang P, Raslan AM, Selden NR. Improving performance by improving education. In: Guillaume DJ, Hunt MA, eds. *Quality and Safety in Neurosurgery*. Cambridge, Massachusetts: Elsevier, Academic Press; 2018:213-224.

Conflict of interest statement: This work was supported by the Di Giovanni Foundation, Montreal English School Board, Colannini Foundation, and Montreal Neurological Institute and Hospital. A. Winkler-Schwartz is supported by a doctoral training grant for applicants with a professional degree issued by the Fonds de recherche du Québec — Santé. R. Del Maestro is the William Feindel Emeritus Professor in Neuro-Oncology at McGill University.

Received 10 December 2018; accepted 6 March 2019

Citation: World Neurosurg. (2019) 127:e230-e235.
https://doi.org/10.1016/j.wneu.2019.03.059

Journal homepage: **www.journals.elsevier.com/world-neurosurgery**

Available online: **www.sciencedirect.com**