Eye (2013) 27, 1269–1274

www.nature.com/eye

CLINICAL STUDY

The development of a virtual reality training programme for ophthalmology: repeatability and reproducibility (part of the International Forum for Ophthalmic Simulation Studies)

GM Saleh[1,2,3], K Theodoraki[2], S Gillan[2], P Sullivan[2], F O'Sullivan[3], B Hussain[2], C Bunce[2] and I Athanasiadis[2]

Abstract

Purpose **To evaluate the variability of performance among novice ophthalmic trainees in a range of repeated tasks using the Eyesi virtual reality (VR) simulator.**

Methods **Eighteen subjects undertook three attempts of five cataract specific and generic three-dimensional tasks: continuous curvilinear capsulorhexis, cracking and chopping, cataract navigation, bimanual cataract training, anti-tremor. Scores for each attempt were out of a maximum of 100 points. A non-parametric test was used to analyse the data, where a *P*-value of <0.05 was considered statistically significant.**

Results **Highly significant differences were found between the scores achieved in the first attempt and that during the second ($P<0.0001$) and third ($P<0.0001$) but not between the second and third attempt ($P=0.65$). There was no significant variability in the overall score between the users ($P=0.1104$) or in the difference between their highest and lowest score ($P=0.3878$). Highly significant differences between tasks were shown both in the overall score ($P=0.0001$) and in the difference between highest and lowest score ($P=0.003$).**

Conclusion **This study, which is the first to quantify reproducibility of performance in entry level trainees using a VR tool, demonstrated significant intra-novice variability. The cohort of subjects performed equally overall in the range of tasks (no inter-novice variability) but each showed that performance varies significantly with the complexity of the task when using this high-fidelity instrument.**

Eye (2013) **27,** 1269–1274; doi:10.1038/eye.2013.166; published online 23 August 2013

Keywords: Cataract; virtual reality; simulator; repeatability; reproducibility; variability

Introduction

Mastery of performance, especially of highly technical tasks such as microsurgery[1,2] and aviation[3] takes years of repeated practice to achieve.[4] In ophthalmology, evidence has emerged for a wider spread of performance in the early stages of training. This has been demonstrated for a range of tasks including cataract surgery,[5] corneal suturing in a wet lab,[6] in oculoplastic surgery,[7] and in motion tracking studies both for phacoemulsification[8] and for skin suturing.[6,9] However, no specific analysis was ever undertaken with regard to the repeatability, reproducibility, or variability among novice trainees but rather these subjects were compared with more senior colleagues.[5,6,8–15] What we therefore do not know is how an individual with minimal experience scores when undertaking the

[1]NIHR Biomedical Research Centre at Moorfields Eye Hospital, NHS Foundation Trust, UCL Institute of Ophthalmology, London, UK

[2]Moorfields Eye Hospital, London, UK

[3]School of Ophthalmology, The London Deanery, London, UK

Correspondence: GM Saleh, NIHR Biomedical Research Centre for Ophthalmology at Moorfields Eye Hospital, NHS Foundation Trust, UCL Institute of Ophthalmology, 162 City Road, London EC1V 2PD, UK
Tel: +44 20 725 33411; Fax: +44 20 7253 4696.
E-mail: George.saleh@moorfields.nhs.uk

Received: 9 January 2013
Accepted in revised form: 4 July 2013
Published online: 23 August 2013

same task on several occasions (variability and reproducibility).

With the emergence of high-fidelity stimulation in the last few years, the potential to deliver modular training in a quantitative manner has now become a possibility with the ability to return a numerical score for analysis. This can therefore be utilised to evaluate a series of metrics and return a standardised score. In this study we employed this facility to examine both intra and inter user repeatability, variability, and reproducibility (differences between an individual's performance and then differences between the individuals themselves). For this purpose, the Eyesi system was used across a range of cataract-specific and generic three-dimensional tasks.

Materials and methods

This prospective study was conducted at Moorfields Eye Hospital with the support of STeLI (Simulation and Technology-enhanced Learning Initiative), the London Deanery School of Ophthalmology and IFOS (International Forum of Ophthalmic Simulation). IFOS is a multinational collaboration set up to investigate and deliver high-fidelity virtual reality (VR) training in ophthalmology. It is administered via a global networked cloud of simulators from participating organisations.

The Eyesi ophthalmosurgical simulator (VR Magic, Manheim, Germany) was used for this study. This comprises a mannequin head with a virtual eye, an operating microscope and a touch screen, all connected to a customised PC. As in a real life-operating situation there are two foot pedals (one for the microscope and one the phacoemulsification), the instruments contain coloured heads from which optical tracking systems convert movements to electrical signals, which are relayed to the simulator after being generated.

Inclusion criteria

All eligible novice ophthalmic trainees with 2 h or less of simulation and intraocular surgical experience were invited to participate.

Exclusion criteria

Novice trainees who did not wish to participate in the study and those with more than 2 h of simulation and intraocular surgical experience were excluded.

IRB (institutional review board)

We certify that all applicable institutional and governmental regulations concerning the ethical use of human volunteers were followed during this research.

Simulator induction

A consultant attending trainer (GS) gave all subjects a standardised simulator induction. Each candidate was also given instructions on the set-up including microscope adjustment, seating, positioning, and foot-pedal use. Each trainee received a personalised account through which all data acquisition was captured. A clear description of all tasks was presented prior to commencement.

VR tasks

Five modules were selected, including one cataract-specific task (capsulorhexis level 1) and four generic three-dimensional tasks (cracking and chopping level 2, cataract navigation level 3, cataract bimanual training level 1, anti-tremor level 2; Figure 1). The ability to differentiate between different levels of expertise (construct validity) has been demonstrated for the capsulorhexis, cracking and chopping, navigation, and anti-tremor modules.[12–14]

The capsulorhexis involves the following steps: injection of viscoelastic into the anterior chamber, flap creation, and capsulorhexis formation. For the cracking and chopping, the trainee is required to pull two spheres simultaneously until they reach a given length illustrated by a change in colour. In cataract navigation, the tip of the instrument is held steady inside a sphere for a set period of time. The cataract bimanual training is a two-handed static task where the junior is asked to hold the tips of two instruments steady at the ends of a cylinder until it changes from green to red. For the anti-tremor module, the subject needs to move a sphere with the tip of the instrument around a cylinder at a fixed speed.

Each one of the tasks was repeated three times to test for repeatability and reliability.

Statistical analysis

Data were analysed using non-parametric tests because of evidence of non-normality. The signed-rank test was used to assess whether scores differed between the first and second attempt and between first and third attempt, and between second and third attempt. The Kruskal–Wallis test was used to assess whether the overall and the range of (highest–lowest) scores differed between individuals and/or between tasks. For all tests, a *P*-value of less than 0.05 was considered significant.

Results

Eighteen subjects were enroled in this study and the results are presented in Tables 1–3. Table 1 outlines the inter-novice results. No significant differences in the scores were demonstrated between the juniors using

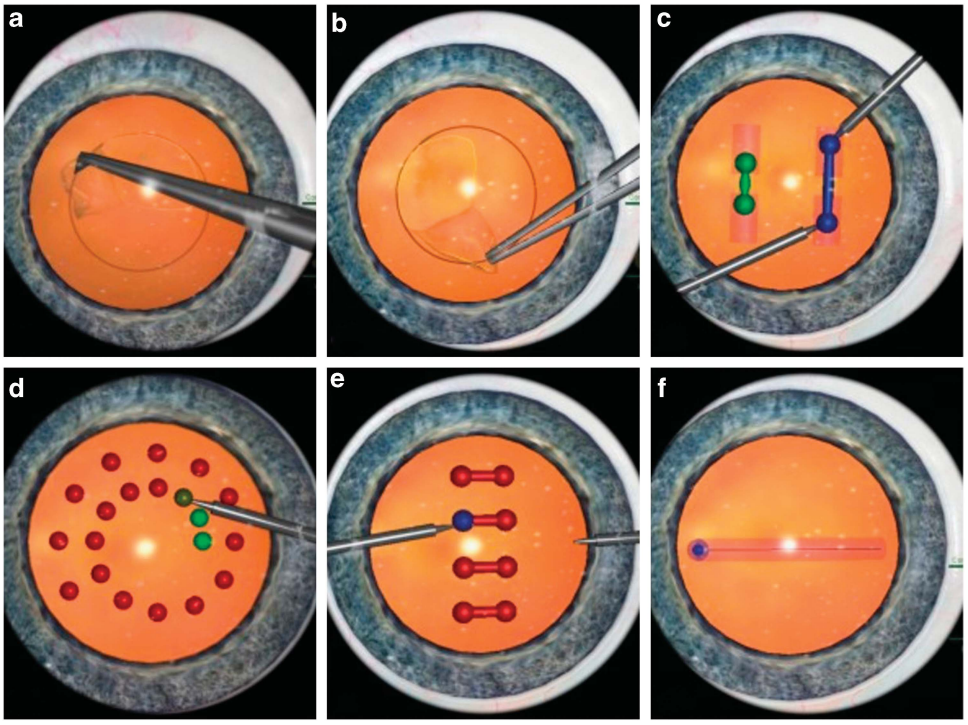

Figure 1 Eyesi simulator tasks. (a, b) Capsulorhexis, (c) cracking and chopping, (d) navigation, (e) bimanual training, (f) anti-tremor.

Table 1 Scores across individuals (inter-novice performance)

Scores in all tasks by trainee				*Difference between highest and lowest score in a single task*			
Trainee	*Lower limit*	*Median*	*Upper limit*	*Trainee*	*Lower limit*	*Median*	*Upper limit*
1	0	32	76	1	35	51	57
2	0	69	70	2	10	18	22
3	0	78	94	3	4	6	61
4	6	64	85	4	21	23	32
5	0	19.5	81	5	0	4	21.5
6	0	72	84	6	6	12	24
7	17	59	90	7	30	46	59
8	11	65	93	8	37	39	65
9	0	34	76	9	8	24	39
10	0	57	86	10	21	28	29
11	0	53.5	75	11	16	47.5	78
12	37	74	88	12	28	36	36
13	0	53	69	13	3	24	42
14	49	77	95	14	30	39	66
15	0	35	73	15	0	21	42
16	0	79	92	16	4	5	6
17	0	62	79				
18	0	0	58				
P-value		0.1104		*P*-value		0.3878	

Table 2 Scores across attempts (intra-novice performance)

2-Scores of all tasks by attempt			
Attempt	*Lower limit*	*Median*	*Upper limit*
1st	0	36.5	
2nd	0	68	83.5
3rd	13	69	83
1st *vs* 2nd attempt		*P*-value <0.0001	
1st *vs* 3rd attempt		*P*-value <0.0001	
2nd *vs* 3rd attempt		*P*-value 0.65	

different tasks ($P=0.1104$). This indicates that in this group of first year trainees there was no significant difference in their ability to execute the tasks. Similar results were found when the difference between the highest and the lowest score was examined ($P=0.3878$).

Trainees' overall performance differed significantly between the first and second attempt ($P<0.0001$) and between the first and third attempt ($P<0.0001$), but not between the second and third attempt ($P=0.65$) as demonstrated in Table 2. This indicates an initial poor reproducibility for the high-fidelity tasks by this group of novice trainees, while a certain level of consistency in scores is achieved between the 2nd and the 3rd attempt ($P=0.65$).

Table 3 summarises the trainees' scores for each task. There were highly significant differences among the results achieved by module ($P<0.0001$). There was a significant difference between the highest and lowest score by task ($P=0.003$). The above shows that the performance varies significantly with the complexity of the task. For example, it is more challenging to perform

Table 3 Scores across tasks

Overall performance by task				*Difference in highest and lowest score in a single task*			
Task	*Lower limit*	*Median*	*Upper limit*	*Task*	*Lower limit*	*Median*	*Upper Limit*
Capsulorhexis 1	0	0	45	Capsulorhexis 1	21	51	61.5
Cracking and Chopping 2	65	89.5	94	Cracking and chopping 2	7	23	38
Cataract navigation 3	42	73.5	78	Cataract navigation 3	23	40.5	47.5
Bimanual training 1	68.5	81	91	Bimanual training 1	6	19.5	28
Anti-tremor 2	0	0	0	Anti-tremor 2	0	1	34
P-value		0.0001		*P*-value		0.0030	

a capsulorhexis where the median and the upper limit scores achieved were 0 and 45, respectively, compared to the navigation module where the median and the upper limit scores were 81 and 91, respectively. Therefore, the simulator appears to have fidelity in being able to deliver different levels of difficulty to cohort of the first year ophthalmic trainees.

Discussion

This is the first study to explicitly examine the repeat performance of novice ophthalmic trainees in a range of intra-ocular tasks. It clearly demonstrates a variable pattern among the different attempts and tasks while all juniors perform roughly similarly. Previous studies in the wet lab environment[7,8] and those scoring actual cataract surgery[5] have also suggested a broader spread of scores in a novice cohort. However, the internal numerical variance and its statistical significance have not been analysed in detail before and this is relevant for the trainers, the individual trainees but also for the design of a training syllabus.

Construct validation on the simulator has been established comparing novice groups with a more senior surgical cohort[12–14] and on average the more experienced performed better. All trainees enroled in this study had previously passed through a competitive selection process and were now in their first year of the London Deanery School of Ophthalmology Programme. So, unlike previous studies they were all at the same early stage of training. Among these juniors, no statistically significant difference (no inter-novice variability) is demonstrated ($P = 0.1104$, Table 1). This result indicates that they all found the tasks equally challenging and had an equally varied performance when repeating the given modules. This trend holds true when the difference among the highest and lowest scores were assessed ($P = 0.3878$). That may imply a more limited usefulness of the simulator in the selection or interview process or at least caution its discriminatory ability in the very inexperienced. Equally there may be substantial potential pitfalls if the simulator is used as an assessment tool at this stage of training, especially in a high-stake scenario such as eligibility of entry into an ophthalmic training programme. All these trainees appear to be starting out on a reasonably even footing. Additionally, even though they were given formative feedback during their assessment, they were unable to modify the consistency of their performance sufficiently and therefore a continuous structured debrief by a supervisor-consultant is clearly of benefit at this stage of training.

There is a clear upward trend of performance with repeated attempts (as shown by the median scores achieved in each of the three repeats, Table 2). Even though there is poor reproducibility when comparing the 1st attempt with 2nd ($P < 0.0001$) and the 3rd ($P < 0.0001$), novice trainees seem to achieve a certain level of competency and consistency on their scores between the 2nd and the 3rd attempt ($P = 0.65$). Importantly therefore, at this earliest stage of training a minimum of three repeats of any given task should be encouraged both for learning and benchmarking. A further 4th attempt could be considered to allow more accurate results and further improvement of performance. Any fewer and spurious results may be obtained. This has significant implications with regard to the necessity of constant clinical supervision in the early stages; competency demonstrated in a single attempt may be misleading as the junior surgeon may be incapable of repeatedly achieving the same performance. The shift of performance and the lack of reproducibility between the first and subsequent tasks also reflects a certain learning curve. Despite simulator's limitations for summative feedback, it is a good training tool as practice along with feedback (formal consultant-led, point deduction for errors occurred and score breakdown) could potentially lead to an improved performance. In the future, studies with increased number of attempts over fewer tasks would allow a 'learning curve' to be established with the potential to identify above- and below-average trainees. Additionally an investigation of a reduction in variability could be explored in relation to increased score as the individuals progress through their training.

Clear evidence of differences in the total scores between the five modules is present in the results ($P = 0.0001$). The more technically challenging tasks,

such as the capsulorhexis, had lower median and upper limit scores (0 and 45, respectively) than relatively easier modules, such as the one-handed static cataract navigation, where these scores were higher (73.5 and 78 respectively). Novice surgeons are prone to a greater number of errors[11] and simulator's internal scoring system deducts points when specific errors are encountered during the procedure. Although arbitrary, it is normalised to mark out of 100 and generated purely objectively having innate accuracy. It is useful, therefore, that the simulator additionally demonstrates a degree of fidelity in this study, by generating relatively appropriate scores based on the difficulty of the exercise. This is helpful for standardising trainee scores, benchmarking progression, and for task selection in future, more structured, ophthalmic simulation programmes.

The data presented show that the simulator would be more useful to monitor performance (formative assessment) rather than to evaluate and quantify overall skills (summative assessment). The highly significant variability demonstrated in juniors' performance offers the first quantitative description of this trend. The performance variance that clearly exists at entry level would be beneficially reduced by repeated simulator practice but further work will be required to confirm the nature of this trend. Given this variability, caution should also be exercised if using this tool for evaluation to help determine competency or entrance at early stages of ophthalmic training.

Summary

What was known before

- Previous reports qualitatively suggested that novice ophthalmic surgeons may have a wider spread of performance in the early stages of their training.
- The emergence of high-fidelity stimulation has allowed novice surgeons to use virtual reality (VR) tools to supplement their live surgical skill acquisition.

What this study adds

- This is the first study to quantify the variability of performance in entry level trainees using a VR tool.
- All novice trainees have appear to have similar performance when repeating a range of tasks.
- When performing the same task over three repeats, all trainees demonstrated significant variability in their performance. Therefore, when designing a relevant syllabus a minimum of three attempts is recommended for benchmarking of performance.
- Trainees' scores in various VR modules showed significant differences reflecting complexity of the modules. This appears to suggest fidelity for the tasks in question.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We acknowledge the Special Trustees of Moorfields Eye Hospital for their unrestricted financial support towards this work. They also acknowledge support from The London Deanery School of Ophthalmology and Simulation and Technology-enhanced Learning Initiative (STeLI), who have supported the development of ophthalmic simulation programmes in the region. George M Saleh acknowledges (a proportion of his) financial support from the Department of Health through the award made by the National Institute for Health Research to Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology for a Biomedical Research Centre.

References

1 Aggarwal R, Black SA, Hance JR, Darzi A, Cheshire NJW. Virtual reality simulation training can improve inexperienced surgeons' endovascular skills. *Eur J Vasc Endovasc Surg* 2006; **31**(6): 588–593.
2 Ikonen TS, Antikainen T, Silvennoinen M, Isojärvi J, Mäkinen E, Scheinin TM. Virtual reality simulator training of laparoscopic cholecystectomies—a systematic review. *Scand J Surg* 2012; **101**(1): 5–12.
3 Longridge T, Bürki-Cohen J, Go TH, Kendra AJ. Simulator fidelity considerations for training and evaluation of today's airline pilots. Proceedings of the 11th international symposium on aviation psychology (Internet). 2001 (cited November 28 2012). p 1–7. Available from https://www.hf.faa.gov/docs/508/docs/VolpeLongridge2001.pdf.
4 Ericsson KA, Prietula MJ, Cokely ET. The making of an expert. *Harvard Business Review* 2007; **85**(7/8): 114.
5 Saleh GM, Gauba V, Mitra A, Litwin AS, Chung AKK, Benjamin L. Objective structured assessment of cataract surgical skill. *Arch. Ophthalmol.* 2007; **125**(3): 363–366.
6 Saleh GM, Voyatzis G, Voyazis Y, Hance J, Ratnasothy J, Darzi A. Evaluating surgical dexterity during corneal suturing. *Arch Ophthalmol* 2006; **124**(9): 1263–1266.
7 Gauba V, Saleh GM, Goel S. Ophthalmic plastic surgical skills assessment tool. *Ophthal Plast Reconstr Surg* 2008; **24**(1): 43–46.
8 Saleh GM, Lindfield D, Sim D, Tsesmetzoglou E, Gauba V, Gartry DS *et al*. Kinematic analysis of surgical dexterity in intraocular surgery. *Arch Ophthalmol* 2009; **127**(6): 758–762.
9 Saleh GM, Gauba V, Sim D, Lindfield D, Borhani M, Ghoussayni S. Motion analysis as a tool for the evaluation of oculoplastic surgical skill: evaluation of oculoplastic surgical skill. *Arch Ophthalmol* 2008; **126**(2): 213–216.
10 Solverson DJ, Mazzoli RA, Raymond WR, Nelson ML, Hansen EA, Torres MF *et al*. Virtual reality simulation in acquiring and differentiating basic ophthalmic microsurgical skills. *Simul Healthc* 2009; **4**(2): 98–103.

11 Gauba V, Tsangaris P, Tossounis C, Mitra A, McLean C, Saleh GM. Human reliability analysis of cataract surgery. *Arch Ophthalmol* 2008; **126**(2): 173–177.
12 Mahr MA, Hodge DO. Construct validity of anterior segment anti-tremor and forceps surgical simulator training modules: attending versus resident surgeon performance. *J Cataract Refract Surg* 2008; **34**(6): 980–985.
13 Privett B, Greenlee E, Rogers G, Oetting TA. Construct validity of a surgical simulator as a valid model for capsulorhexis training. *J Cataract Refract Surg* 2010; **36**(11): 1835–1838.
14 Selvander M, Asman P. Cataract surgeons outperform medical students in Eyesi virtual reality cataract surgery: evidence for construct validity. *Acta Ophthalmol* 2013; **91**(5): 469–474.
15 Spiteri A, Aggarwal R, Kersey T, Benjamin L, Darzi A, Bloom P. Phacoemulsification skills training and assessment. *Br J Ophthalmol* 2010; **94**(5): 536–541.