High correlation between performance on a virtual-reality simulator and real-life cataract surgery

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Acta Ophthalmologica</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>ACTA-16-02-0219.R2</td>
</tr>
<tr>
<td>Wiley - Manuscript type:</td>
<td>Original Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>01-Aug-2016</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Thomsen, Ann Sofia; Rigshospitalet - Glostrup, Department of Ophthalmology; Rigshospitalet, Copenhagen Academy for Medical Education and Simulation
Smith, Philip; University of Surrey, Department of Computing
Subhi, Yousif; Rigshospitalet, Copenhagen Academy for Medical Education and Simulation
la Cour, Morten; Glostrup Hospital, Department of Ophthalmology
Tang, Lilian; University of Surrey, Department of Computing
Saleh, George; University of Surrey, Department of Computing; Moorfields Eye Hospital NHS Foundation Trust, Department of Education; Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology, The National Institute for Health Research
Konge, Lars; Rigshospitalet, Copenhagen Academy for Medical Education and Simulation |
| Keywords: | assessment, cataract surgery, virtual-reality simulation, motion-tracking, validity |
High correlation between performance on a virtual-reality simulator and real-life cataract surgery

Ann Sofia Skou Thomsen, MD¹²
Phillip Smith, BSci³
Yousif Subhi, MD²
Morten la Cour, Professor, DMSc, MD¹
Lilian Tang, BEng, MEng, PhD³
George M. Saleh, FRCSEd, FRCOphth³⁴⁵
Lars Konge, PhD, MD²

¹ Department of Ophthalmology, Rigshospitalet - Glostrup, Copenhagen, Denmark.
² Copenhagen Academy for Medical Education and Simulation, Rigshospitalet, Copenhagen, Denmark.
³ Department of Computing, University of Surrey, Guildford, England.
⁵ The National Institute for Health Research, Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology, Moorfields Eye Hospital, London, England.

Corresponding author: Ann Sofia Skou Thomsen, Department of Ophthalmology, Glostrup University Hospital, Ndr. Ringvej 57, DK-2600 Glostrup, Denmark. Email: skouthomsen@dadlnet.dk. Telephone no: +45 38634700. Fax no. +45 38634669.
Abstract:

**Purpose:** To investigate the correlation in performance of cataract surgery between a virtual-reality simulator and real-life surgery using two validated, objective assessment tools.

**Methods:** Cataract surgeons with varying levels of experience were included in the study. All participants performed and video-recorded three standard cataract surgeries before completing a proficiency-based test on the EyeSi virtual-reality simulator. Standard cataract surgeries were defined as: 1) surgery performed under local anaesthesia, 2) patient age > 60 years, and 3) visual acuity > 1/60 pre-operatively. A motion-tracking score was calculated by multiplying average path length and average number of movements from the three real-life surgical videos of full procedures. The EyeSi test consisted of five abstract and two procedural modules: intracapsular navigation, antitremor training, intracapsular antitremor training, forceps training, bimanual training, capsulorhexis, and phaco divide and conquer.

**Results:** Eleven surgeons were enrolled. After a designated warm-up period, the proficiency-based test on the EyeSi simulator was strongly correlated to real-life performance measured by motion-tracking software of cataract surgical videos with a Pearson correlation coefficient of -0.70 (p=0.017).

**Conclusion:** Performance on the EyeSi simulator is significantly and highly correlated to real-life surgical performance. However, it is recommended that performance assessments are made using multiple data sources.
Key Words: Assessment, cataract surgery, virtual-reality simulation, motion-tracking, validity.
Introduction:

Increased focus on proficiency-based training (i.e. training to a pre-defined competency level) in surgical disciplines has led to a demand for objective assessment tools (O’Day 2007, Oetting 2009). Several assessment tools for cataract surgical skills have been developed (Gensheimer, Soh & Khalifa 2013, Puri & Sikder 2014), but the majority are based on clinical performance evaluations by human raters which make them labour intensive and potentially prone to bias (Konge et al. 2012, Williams, Klamen & McGaghie 2003). Virtual-reality simulator metrics may have the ability to provide more objective measurements of surgical skills, as no rater is included in the assessment process. The EyeSi simulator is by far the most investigated virtual-reality simulator in ophthalmic surgery, including cataract surgery (Thomsen et al. 2015) but thus far, we have no evidence supporting the direct translation between the automated, summative assessment provided by the simulator, i.e. the simulation-based performance score, and real-life surgical performance.

Several studies have looked into the validity of the simulator metrics in order to investigate the use of the simulator as an assessment tool (Le, Adatia & Lam 2011, Selvander & Åsman 2012, Thomsen et al. 2015). Evidence of validity has primarily been demonstrated by the ability to differentiate between novices and experienced surgeons, i.e. using the experience level of participants as a proxy parameter for surgical competency. Confounders may therefore be introduced, and the ability to differentiate between these groups is indeed a necessary but not sufficient feature of an assessment tool (Cook 2015). A more direct method of exploring the relevance of the EyeSi is to compare the simulator metrics to real-life surgical performance. This has been done for training duration by Daly et al. who show a significant correlation between time
to complete an EyeSi training course and subsequent performance in the operating room.

However, it has not been done for the summative assessment score provided by the simulator, including various outcomes such as target achievement, instrument handling and tissue treatment in addition to time.

For the objective assessment of surgical skills in a clinical setting, motion analysis has been validated for number of hand movements, path length, and time taken for corneal suturing, oculoplastic surgery, and cataract surgery (Saleh et al. 2008, Saleh et al. 2006, Smith et al. 2013). These findings are in accordance with the Fitts-Posner Three Stage Theory of motor skill acquisition: performance becomes more fluid and with fewer interruptions as the trainee becomes more experienced, leading to fewer hand movements and decreased path length (Reznick & MacRae 2006).

The purpose of this study was to investigate the correlation between performance on a virtual-reality simulator (EyeSi performance score) and motion-tracking metrics from real-life cataract surgery. We hypothesize a good correlation between these two objective assessment tools – both designed to measure surgical performance – in all ends of the performance scale.
Methods:

This study was designed as a cross-sectional study. Cataract surgeons with different surgical experience levels employed at an ophthalmology department or a private clinic in Denmark were invited to participate in the study. Only surgeons operating independently were included. Stereoacuity measurement using the TNO test (Laméris Ootech BV, 16th edition) was also a criterion of inclusion. Exclusion criteria were participants with more than 2 hours of virtual-reality simulator experience within the last 6 months, and poor technical quality of surgical videos. This study adhered to the tenets of the Declaration of Helsinki. All participants signed informed consent and completed a questionnaire regarding previous surgical and simulator experience prior to enrolment in the study. The Ethics Committee of the Capital Region of Denmark ruled that approval was not required for this study (protocol no. H-6-2014-011).

The period of enrolment was from April 15 to November 15 2014. All surgeons who responded to the invitation in this 7-month period were included in the study. We collected videos of the last three standard cataract surgeries consecutively performed by each surgeon prior to their simulation session. Standard cataract surgeries were defined as: 1) surgery performed in local anaesthesia, 2) patient age > 60 years, and 3) visual acuity > 1/60 pre-operative. The videos of full procedures were analysed in the period December 01 2014 to April 30 2015, using previously validated motion-tracking software (“PhacoTracking”) parameters, including total path length and number of movements (Smith et al. 2013). See Figure 1 for a demonstration of the motion-tracking software. The cataract surgeries were performed at various ophthalmology departments or private clinics in Denmark.
All participants then completed a previously validated proficiency-based test of cataract surgical skills on the EyeSi™ simulator (VRmagic, version 2.8.10) (Thomsen et al. 2015). A maximum of two weeks were accepted between video recordings and the following simulation session. The test consisted of five abstract modules and two procedural modules: intracapsular navigation (level 2 of 3), antitremor training (level 4 of 7), intracapsular antitremor training (level 2 of 5), forceps training (level 4 of 4), bimanual training (level 5 of 5), capsulorhexis (level 1 of 3; weak zonules; no initial tear), and phaco divide and conquer (level 5 of 8). The difficulty level of each module was based on results from a pilot study and previous validation studies on the EyeSi simulator (Thomsen et al. 2015). The test session were preceded by a 10-minutes standardized introduction to the simulator and two warm-up sessions (corresponding to approximately 1-hour). The simulator sessions were performed at the Simulation Centre at Rigshospitalet, Capital Region of Denmark.

The primary outcomes for this study were simulation-based test score and motion-tracking score for each surgeon. The simulator software assessed between 21 and 33 different outcomes for each module from four different domains: Target achievement, efficiency, instrument handling, and tissue treatment. A total score between zero and 700 points was provided based on this assessment. The motion-tracking score was calculated by average path length multiplied by average number of movements based on their three full live cataract procedures captured on videos. Based on Messick’s modern validity framework, we investigated “relations to other variables”, i.e. the correlation of the EyeSi test score to an external, independent measure (Messick 1990).
SPSS software version 19.0 (SPSS, Inc., Chicago, IL) was used for statistical analysis. Linear regression analysis was performed to investigate the correlation between simulator metrics and motion-tracking score. Pearson correlation coefficient was calculated between the test score and the mean motion-tracking score. A higher test score on the EyeSi simulator was considered better (Thomsen et al. 2015) while a lower motion-tracking score reflected superior performance (Saleh et al. 2008). A positive Pearson correlation indicated that both variables increased or decreased together, whereas a Pearson correlation coefficient of -1 demonstrated perfect negative correlation, i.e. linear correlation between high test score and low motion-tracking score, and vice versa.
Results:

Eleven cataract surgeons from six different ophthalmology departments and private clinics in Denmark were included in the study. Two surgeons were excluded from the study due to poor technical quality of surgical videos. The included surgeons were between 38 and 58 years old (median 41), eight were males (73 %), and ten were right-handed (91 %). They had a stereoacuity between 30 and 240 seconds of arc (median 30) measured by TNO test. The number of cataract surgeries performed by each surgeon varied from two to 24,200 (mean 3,656), most surgeons having performed less than 1,000 surgeries.

There was a range of two to eleven days (mean 5) between recording of videos and the simulator session, and none of the surgeons performed cataract surgery or other types of intraocular surgeries during this period. All eligible surgeons had limited experience with a virtual-reality simulator prior to enrolment in this study and no surgeons were excluded from the study due to excess experience with a virtual-reality simulator: Three surgeons had spent approximately 1 hour on the EyeSi simulator 6 months before, and two surgeons had spent 1 hour and 7 hours on the EyeSi simulator two years prior to this study. Most of the included surgeons did not have any previous experience with a virtual-reality simulator.

Table 1 shows performance details for each surgeon. We found a significant association between simulation-based test score and motion-tracking score (p-value=0.017, adjusted $R^2=0.43$), see Figure 2. In particular the expected change in motion-tracking score due to a ten unit change in simulation-based test score was estimated to -23,220 with a 95% confidence interval of (-41,240, -5,210).
Discussion:

In this study we found that a proficiency-based test on the EyeSi simulator showed significant correlation to real-life performance, measured by motion-tracking software, of cataract surgical videos. To our knowledge, this is the first study to show a direct correlation between different objective measures of cataract surgical performance in addition to time, respectively in a virtual-reality environment and in the operating room. This is an important finding with regards to simulator use as a training and assessment tool and has direct relevance to actual surgery.

It has previously been shown that a proficiency-based test on the EyeSi simulator provides a reliable and valid assessment of cataract surgical skills when using the experience of the surgeons as a proxy measure of clinical proficiency (Thomsen et al. 2015). One of the strengths of the present study is that the simulator score correlates well with real surgical performance in the operating room without using experience level as a proxy parameter for proficiency. Including experience as a separate variable may result in incorrect estimates due to confounding factors, because the variables directly affecting cataract surgical experience level has not been mapped. For example, it is unclear whether other intra-ocular surgeries should be included in the model; One of the cataract surgeons included in this study had completed > 300 vitreoretinal surgeries at the time of inclusion which may have added to the proficiency level of this particular surgeon, and might explain why experience level is not strongly correlated with motion tracking score or simulator score in this cohort (Figure 2).

Adding several supplementary variables to the regression model would probably account for some of the unexplained variance in the presented model. Time is an objective quantifiable outcome
measure, which has previously been shown to correlate with surgical proficiency (Daly et al. 2013, Saleh et al. 2009, Smith et al. 2013) and is included in the simulator score. However, small sample size did not allow for us to conduct a multiple regression analysis including several variables, e.g. time to complete cataract surgery, or stereopsis, in the regression model.

It has been difficult to recruit participants as a result of a limited number of cataract surgeons in Denmark. However, despite a small sample size we found a statistically significant correlation and believe that future, bigger studies would corroborate these findings. Another limitation to our study is the risk of selection bias and confounding due to surgical comorbidities. The surgeons recorded and collected the surgical videos themselves, which could have led to unequal difficulty of the included cases.

The rather big variance in the standard error of the estimate makes the predictive value of the simulator only relevant in conjunction with other assessments of proficiency. It is generally acknowledged that multiple data source assessments are the only reliable method for measuring surgical competency. This is in accordance with existing guidelines on this area in ophthalmology: Lee et al. recommend at least two different tools to assess surgical competency (Lee & Carter 2004).

When considering the EyeSi simulator as an assessment tool, we also need to take a warm-up effect of the simulator into account, i.e. performance on the simulator may be influenced by getting to know the technical features of the simulator (Ward, Williams & Hancock 2006). From previous studies, we know that even very experienced surgeons improve their performance score when being assessed on the simulator initially (Thomsen et al. 2015). This effect may affect the
initial performance score differently for all trainees, and subsequently lead to unfounded decisions based on the assessments. One approach to minimize this type of bias is to deliver a 1-hour warm-up on the simulator prior to simulation-based assessment as done by Schijven et al (Schijven & Jakimowicz 2004).

Previous work has shown that repeated practice on the simulator improves the scores it measures (Saleh et al. 2013). The correlation reported here between simulator scores and real-life performance could well indicate that an increase in score on the simulator improves operating room performance but future studies are needed to prove this association.

There are several benefits of using virtual-reality simulators in surgical education besides unbiased assessment of surgical skills, including the fact it is less time intensive for educators as the simulator also provides formative feedback. A limitation to using the EyeSi simulator as an assessment tool in all residency programs is cost. However, Lowry et al found that purchase of the EyeSi simulator was related to a wide range of cost reductions between residency programs with number of residents being a major predictor (Lowry, Porco & Naseri 2013). This in spite of the fact that only a decrease in operating time after training on the EyeSi simulator was included in the equation. A decrease in time commitment for faculty members / educators may be added to the estimate in future analysis. A limitation to “PhacoTracking” as an assessment tool is the requirement of a centralized image in the surgical videos, which may be difficult to achieve for the less experienced surgeon. If the image is not centralized, the computer software will detect less movements and shorter path length. However, corrections can be made to account for these potential errors.
Training of surgeons to achieve competency is an important and sometimes difficult task for residency programs. Time-directed training does not ensure competent surgeons, thus the shift towards proficiency-based training (O’Day 2007). In order to meet the need for assessment tools, we have previously developed a proficiency-based test on the EyeSi simulator with evidence of validity and a reliability coefficient of 0.76 (considered sufficient for moderate stakes assessments) (Thomsen et al. 2015). The results of the present study show that test performance on the EyeSi simulator is highly correlated to real-life surgical performance measured by motion-tracking software, and underpins the use of the EyeSi simulator as an objective assessment tool in ophthalmology. We have to bear in mind, that the presented results are related to the specific test applied, and that a warm-up period is necessary before true competency can be assessed. Future studies should investigate further predictors of surgical proficiency in the operating room in addition to the costs and effects of implementing a virtual-reality simulator in residency programs in order to further clarify its warranted role in training and objective assessment of surgical skills.
Acknowledgements:

Conflict of Interest Disclosures: None reported.

Funding: The study was partly funded by Fight for Sight Denmark and Synoptik Foundation. The author (GS) acknowledges the Department of Health through the award made by the National Institute for Health Research (NIHR) Biomedical Research Centre based at Moorfields Eye Hospital NHS Foundation Trust and the UCL Institute of Ophthalmology. The funding organizations had no role in the design or conduct of this research.

Authors’ contributions: ASST, LK and GMS conceived the design of the study. All data collection was done by ASST. YS administered the videos. The PhacoTracking analysis was performed at Department of Computing, University of Surrey, Guildford, England by PS and LT. ASST, LK and MLC did the final data analysis and/or data interpretation. ASST wrote first draft of the paper and all authors revised the work critically, and approved the final version of the manuscript. The author (ASST) had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
References:


Figure legends:

Figure 1: Demonstration of motion-tracking during (a) incision, (b) Continuous Curvilinear Capsulorhexis (CCC) and (c) phacoemulsification; the coloured dots are points on the instrument where their motion is being measured automatically.

Figure 2: Scatterplot and regression line of the correlation between EyeSi test score and video-based motion-tracking score; the different markers demonstrate different experience levels of the surgeons.
Table 1: EyeSi simulator test score and average motion-tracking outcomes for each individual surgeon and the overall mean (and standard deviation).

*Experience level: 1 = 0-100 surgeries, 2 = 101-500 surgeries, 3 = 501-1000 surgeries, 4 = >1000 surgeries.

<table>
<thead>
<tr>
<th>Surgeon number</th>
<th>Experience level*</th>
<th>Steropsis (TNO)</th>
<th>Average path length</th>
<th>Average number of movements</th>
<th>Motion-tracking score</th>
<th>EyeSi test score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1x</td>
<td>1</td>
<td>30</td>
<td>264.31</td>
<td>2394.67</td>
<td>632,935</td>
<td>485</td>
</tr>
<tr>
<td>2x</td>
<td>1</td>
<td>30</td>
<td>198.85</td>
<td>752.00</td>
<td>149,535</td>
<td>611</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>30</td>
<td>163.42</td>
<td>1029.33</td>
<td>168,212</td>
<td>547</td>
</tr>
<tr>
<td>4x</td>
<td>2</td>
<td>60</td>
<td>238.96</td>
<td>2730.67</td>
<td>652,521</td>
<td>504</td>
</tr>
<tr>
<td>5x</td>
<td>2</td>
<td>60</td>
<td>117.58</td>
<td>1232.00</td>
<td>144,859</td>
<td>661</td>
</tr>
<tr>
<td>6x</td>
<td>3</td>
<td>30</td>
<td>181.19</td>
<td>200.00</td>
<td>362,380</td>
<td>522</td>
</tr>
<tr>
<td>7x</td>
<td>3</td>
<td>30</td>
<td>167.79</td>
<td>1920.00</td>
<td>322,156</td>
<td>616</td>
</tr>
<tr>
<td>8x</td>
<td>3</td>
<td>60</td>
<td>138.61</td>
<td>1088.00</td>
<td>150,808</td>
<td>518</td>
</tr>
<tr>
<td>9x</td>
<td>3</td>
<td>30</td>
<td>135.82</td>
<td>256.00</td>
<td>34,770</td>
<td>631</td>
</tr>
<tr>
<td>10x</td>
<td>4</td>
<td>240</td>
<td>166.10</td>
<td>1495.37</td>
<td>248,381</td>
<td>621</td>
</tr>
<tr>
<td>11x</td>
<td>4</td>
<td>60</td>
<td>118.60</td>
<td>1488.00</td>
<td>176,477</td>
<td>609</td>
</tr>
<tr>
<td>Mean (SD)</td>
<td></td>
<td></td>
<td>172 (47)</td>
<td>1,490 (729)</td>
<td>276,640 (202,015)</td>
<td>575 (61)</td>
</tr>
</tbody>
</table>
(a) Incision

(b) CCC

(c) Phacoemulsification