PhD thesis
Ann Sofia Skou Thomsen

Intra-ocular surgery
– assessment and transfer of skills using a virtual-reality simulator

Academic advisors: Morten la Cour, Lars Konge

WILEY
Preface

This thesis has been submitted to the Graduate School of The Faculty of Health and Medical Sciences, University of Copenhagen. Submitted: 09/02/2017

Academic advisors:

Morten la Cour, Professor, MD, DMSc
Department of Ophthalmology
Rigshospitalet – Glostrup, University of Copenhagen, Denmark

Lars Konge, Professor, MD, PhD
Copenhagen Academy for Medical Education and Simulation
Rigshospitalet, University of Copenhagen, Denmark

Assessment committee:

Jette Led Sørensen, Associate Professor, MD, MMed, PhD (Chairman)
Department of Obstetrics and Gynaecology
Rigshospitalet, University of Copenhagen, Denmark

E. Matthew Ritter, Associate Professor of Surgery, MD, FACS
Department of Surgery, Uniformed Services University of the Health Sciences & the Walter Reed National Military Medical Center, Bethesda, US

Jesper Hjortdal, Professor, MD, DMSc
Department of Ophthalmology
Aarhus University Hospital, Aarhus University, Denmark
Contents

ACKNOWLEDGEMENTS  i
Conflicts of Interest  i

SUMMARY  1

PREFACE  3
Definitions  3

INTRODUCTION  3

BACKGROUND  4
Intraocular surgery  4
Cataract surgery  4
Vitreoretinal surgery (VRS)  4

Acquisition of technical skills  4
Motor skill theories  5
Characterisation of expertise  5
Efficient instructional strategies  5
Proficiency-based training  5

Assessment of technical skills  5
Validity evidence  6
Reliability  7
Benchmark criteria (standard setting)  7

Outcome measures – efficacy studies  7
Transfer of skills from one modality to another (virtual-reality to OR)  8
Transfer of skills from one procedure to another  8

Surgical simulators in ophthalmology  9
Virtual-reality simulators  9

HYPOTHESIS  9

RESEARCH OBJECTIVES  9

SUMMARY OF STUDIES  10

STUDY I: Update on simulation-based surgical training and assessment in ophthalmology – a systematic review  10
STUDY II: Simulation-based certification for cataract surgery  11
STUDY III: High correlation between performance on a virtual-reality simulator and real-life cataract surgery  11
STUDY IV: Operating room performance improves after proficiency-based virtual reality cataract training  12
STUDY V: Is there inter-procedural transfer of skills in intraocular surgery? – a randomized controlled trial  12

DISCUSSION  13
Automated assessment of technical skills  14
Specificity of training interventions  15
Intermediate level surgeons and virtual-reality training  16
Guidelines for simulation-based training in intraocular surgery  16
Implications for future research  16

CONCLUSION AND PERSPECTIVES  17

REFERENCES  17
Acknowledgements

I present my deepest gratitude to both of my supervisors, who have given me the opportunity to work on this project with the perfect combination of independence and support:

**Professor Morten la Cour, MD, DMSci,** for his inspiring mind, and immense knowledge not only in research.

**Professor Lars Konge, MD, PhD,** for his insightful comments and encouragement, but also for being a personal mentor.

My sincere thanks also go to:

**Jens Folke Kielgaard, MD, PhD,** for sharing his great knowledge of research with me, for his motivational talks, and for his always friendly attitude. Also a special thanks to his family.

Furthermore, I would like to thank:

**Participants** in my studies: Without their participation, this project would never have been done.

**PhD students, medical students, and employees** at the Department of Ophthalmology and at CAMES, Rigshospitalet.

Participants at the Wilson Center, University of Toronto, for an inspiring research stay: A special thanks to PhD Mahan Kalasegaram and PhD Ryan Brydges.

**Co-authors** for valuable feedback and discussions.

A special thanks to:

George M. Saleh, FRCSEd, FRCOphth at Moorfields Eye Hospital, London, for sharing interest in this small research area in ophthalmology.

Tobias Todsen and Martin Tolsgaard, MD, PhD for enlightening me the first glance of research in this area.

Mikael Henriksen, MD and Jacob Melchior, MD, PhD at CAMES for many interesting discussions and good laughs.

My office mates Nina Sorensen and Mette Bagger with whom I have shared an office from the beginning.

Jeffrey Chung and David Rojas for some interesting discussions during my research stay in Toronto, Canada.

My family and friends. Especially to Henrik and Theodor.

I would like to acknowledge the Department of Ophthalmology, Rigshospitalet – Glostrup, and Copenhagen Academy for Medical Education and Simulation, Rigshospitalet, for providing premises and administrative assistance throughout the Ph.D. project. Also a special thanks to Kirsten G. Engel MD, researcher at CAMES, for proofreading this thesis.

I would like to thank the funding organizations, whose support has made this project possible:

- Fight for Sight Denmark (Værn om Synet)
- The Synoptik Foundation (Synoptikfonden)
- The Research Council of Rigshospitalet
- Graduate engineer Frode V. Nyegaard and wife’s Grant
- The Capital Region of Denmark
- The Johannes Fogs Grant
- The Hede Nielsen Family Foundation
- The Danielsen Foundation
- Danish Ophthalmological Society
- University of Copenhagen, Faculty of Health Sciences

**Conflicts of Interest**

The author has no conflicts of interest to declare.
Intraocular surgery – assessment and transfer of skills using a virtual-reality simulator

Ann Sofia Skou Thomsen

Department of Ophthalmology, Rigshospitalet – Glostrup, University of Copenhagen, Copenhagen, Denmark

Summary

Virtual-reality training of surgical skills is a rapidly expanding area of research. The implementation of virtual-reality training of surgeons has shown to have implications for patient safety, but many questions remain unanswered. This Ph.D. project began with a systematic review of the existing knowledge in the field. Thereby, knowledge gaps were identified, and the following studies were planned accordingly. This thesis covers research conducted in three related domains: (1) how best to assess surgical trainees’ competence; (2) how to optimize skill acquisition using simulation-based training modalities, including virtual-reality simulation; and (3) which factors influence skill transfer between different environments and different procedures.

In the systematic review, we aim to identify the evidence behind the use of simulation-based models in ophthalmology. Due to heterogeneity of the included trials, a qualitative analysis is conducted. We conclude that limited data are available to support the use of simulation-based models for assessment purposes. Even though numerous studies have investigated the use of the automated assessment metrics provided by the EyeSi virtual-reality simulator, validity evidence has not been well established. Efficacy trials show a tendency towards improved surgical performance, including procedural time, and a decrease in complication rates after implementation of virtual-reality training. However, data are limited, and the results are inconsistent so no final conclusions can be made.

The second study investigates how best to assess surgical trainees’ competence using virtual-reality simulation and an evidence-based performance test of cataract surgical skills on the EyeSi simulator was developed. A total of 42 participants were included, and modules showing discriminative ability between novices and experienced cataract surgeons were included in the final performance test. A benchmark criterion was determined and may be used for future implementation of simulation-based training for novices in cataract surgery. An additional validity study investigates the correlation between virtual-reality performance and motion-tracking metrics from real-life cataract surgeries. Eleven cataract surgeons with different experience levels were included in a national, multicentre study. In this study, we demonstrate that performance on the EyeSi simulator is highly correlated with real-life surgical performance, and may supplement clinical assessments of cataract surgical skills. However, motion-tracking metrics are associated with high levels of interindividual variance, and multiple data sources are still recommended when evaluating surgical skills.

In the fourth study, we examine the impact of proficiency-based training, during which the trainee continues training until passing a predefined proficiency criterion. This approach differs from the traditional time-based or repetition-based training programmes. The study was conducted as a national, multicentre study, involving 22 surgeons with different levels of experience who performed three video-recorded cataract surgeries before and after completing a proficiency-based training programme on the EyeSi simulator. The real-life performance evaluations from three masked raters demonstrate a significant effect of virtual-reality training for novice surgeons, as well as intermediate level surgeons, who had performed up to 75 independent operations before completing the simulation-based training. Thus, improvements in surgical skills resulting from proficiency-based training in cataract surgery seem to be transferable from a simulated setting to the operating room.

Lastly, we wanted to investigate the specificity of skill acquisition in intraocular surgery – specifically, the potential for skill transfer from
The thesis is based on the following original papers:


Introduction

Every day in the healthcare system, adverse events occur in the context of surgical interventions and may result in suboptimal and sometimes devastating outcomes for patients (Makary & Daniel 2016). Risk factors include a variety of patient factors such as age, comorbidities and smoking status, as well as a surgeon factor: more experienced surgeons generally have superior surgical skills as compared to less experienced surgeons, and consequently, their surgical outcomes are better and associated with lower complication rates (Johnston et al. 2010; Ti et al. 2014; Day et al. 2015; Mahmud et al. 2015).

Nevertheless, we need new surgeons. Cataract is the world’s leading cause of blindness and impaired vision, and presently, the only effective treatment is surgical extraction. Today, cataract surgery is one of the most commonly performed surgical procedures in Western countries. The need for ophthalmic surgeons is expected to rise in the future due to an increasing geriatric population, as well as the inevitable ageing of our existing surgeons (Etzioni et al. 2003; Behndig et al. 2011; Kessel 2011; Gollogly et al. 2013).

While some risk factors associated with adverse events and suboptimal patient-related outcomes are nonmodifiable or inevitable, evidence suggests...
that simulation-based training of surgeons has the potential to improve surgical outcomes (Stefanidis et al. 2014). Simulation can be described as ‘something that is made to look, feel, or behave like something else especially so that it can be studied or used to train people’ (Merriam-Webster definition). Simulation-based training – comprising a wide range of simulation models from pig eyes to highly sophisticated virtual-reality models – enables a safe training environment without any associated patient risk.

Virtual-reality simulation, also called technology-enhanced simulation, is the use of interactive computer software and hardware to replicate a real environment. Specifically of interest for this thesis, virtual-reality simulation has been associated with moderate to large effect sizes with respect to surgical skill acquisition and patient-related outcomes for a variety of procedures. Previous studies indicate that surgical skills often improve with virtual-reality training, and the acquired skills may lead to an optimized performance in the operating room, and in turn, contribute to improved patient-related outcomes (Cook et al. 2013c). Thus, virtual-reality training has the potential to improve patient safety by improving the surgeon factor in surgical interventions.

Another major benefit of using virtual-reality simulation is that it enables automated – and thereby objective – assessment of performance, providing the trainee and other key persons with independent feedback. For decades, the assessment and selection of future surgeons have been dependent exclusively on the subjective opinions of senior colleagues (Darzi et al. 1999; Muttuvelu & Andersen 2016). The introduction of virtual-reality simulators provides an opportunity for advancement in the assessment of surgical skills. In summary, virtual-reality simulation consists of two features: (1) a training environment; and (2) a skills assessment component.

**Background**

In ophthalmology, the use of simulation-based training and assessment has increased significantly during the last decade (McCannel 2015). Similarly, the body of research evidence has been growing, but the scientific evidence for using simulation-based methods in ophthalmology is still relatively limited. In laparoscopic surgery, there is a larger quantity of evidence supporting different aspects of simulation-based training (Zendejas et al. 2013). Yet, substantial differences in surgical technique make it difficult to transfer the results directly to ophthalmic surgery.

### Intraocular surgery

Intraocular surgery incorporates all procedures performed within the eye, an organ of approximately 24 mm in diameter. It requires an operating microscope, which is controlled by a foot pedal and enables stereoscopic vision. Some of the instruments also include an operating machine (e.g. one that provides a vacuum and/or ultrasonic energy), which is also navigated via a foot pedal (instrument pedal). Thus, all four extremities are often used simultaneously during parts of an intraocular procedure.

The fact that the procedures involve microsurgery is indeed one of the challenges in ophthalmic surgical training. Visuospatial awareness and/or stereoscopic vision play an important role (Nibourg et al. 2015), small incorrect movements can cause injury, highly specialized instruments are used, and hand-foot co-ordination, in addition to eye-hand co-ordination, is essential. Additionally, supervision is difficult as only one can perform the procedure at a time, and if a change of surgeons is needed, the inserted instruments must be removed from the eye and surgeons exchange seats (followed by adjustment of microscope and chair). These factors make the traditional apprenticeship method less suitable.

Another important feature of intraocular surgery is the traditionally long learning curves for novice surgeons. For cataract surgery, surgical competency improves significantly after the initial 75–80 cases (Tarbet et al. 1995; Randleman et al. 2007; Taravella et al. 2011). Even for surgeons performing more than 500 surgeries annually, the risk for adverse events is generally higher when compared to surgeons performing more than 1000 surgeries annually (Bell et al. 2007).

### Cataract surgery

Cataract surgery is most commonly performed using phacoemulsification, but other techniques include manual small incision cataract surgery (MSICS) and extracapsular cataract extraction (ECCE). All techniques include removal of the opaque lens, which is causing impaired vision and/or visual disturbances of different degrees.

Phacoemulsification consists of six major steps: (1) incisions into the anterior chamber; (2) continuous curvilinear capsulorrhexis (CCC); creation of a circular hole in the anterior capsule; (3) hydrodissection; dissecting the lens capsule from the lens matter; (4) phacoemulsification; removal of the lens matter from the capsule using vacuum and/or ultrasonic energy; (5) irrigation and aspiration (I/A), aspiration of residual soft lens matter, and finally; (6) intraocular lens insertion.

Step 4 (phacoemulsification) is performed using various chopping techniques or ‘divide and conquer’ – the latter being the most common approach, especially for novice surgeons (Alexander et al. 2012; Sorensen et al. 2012). The entire procedure is performed through incision points <4 mm in size, and rotation around the incisional axis is crucial to avoid trauma to the surrounding tissue. Several of the steps are performed bimanually. The steps perceived as most difficult by surgical trainees and with lowest completion rates are phacoemulsification (here: ‘divide and conquer’) and capsulorrhexis, followed by I/A and lens insertion (Dooley & O’Brien 2006).

Serious adverse events include posterior capsule tear and vitreous loss, which are strongly associated with substantial visual loss due to an increased risk of other serious adverse events such as retinal detachment and endophthalmitis (Ti et al. 2014; Day et al. 2015). Randleman et al. (2007) report that vitreous loss occurs in 5.1% of cases for novice surgeons and decreases to 1.9% after the surgeon has completed 80 cases.

### Vitreoretinal surgery (VRS)

Vitreoretinal surgery, another type of intraocular surgery, comprises a variety of procedures performed in the posterior part of the eye and includes retinal detachment repair, macular hole surgery and peeling of epiretinal...
membranes (ERM) among others. Similar to cataract surgery, fine motor skills are critical, in addition to procedural planning and continuous integration of perceptual information. The surgical instruments, including their lengths, are different from cataract surgery, but the concept of a rotation axis through the incisional point is similar. In most countries, vitreoretinal surgery is typically performed by ophthalmologists subspecializing in this area of surgery, whereas residents often perform minor procedures, such as panretinal photocoagulation and intravitreal injections (Shah et al. 2009).

Acquisition of technical skills

Repeated exposure in the clinical setting has been the traditional way of teaching technical skills in the healthcare system. However, numerous factors including patient safety concerns, increased efficiency demands and the constant development of new technology and new treatment modalities, coupled with an increase in patient numbers, have lead to the development of alternative teaching methods. In the following sections, different theoretical approaches for the teaching of technical skills will be reviewed. This is followed by a discussion of more practical approaches, where evidence-based training strategies will be described.

Motor skill theories

The acquisition of technical skills, also called motor skills, has been described by several theoretical models. One such model is the Fitts–Posner Three-Stage Theory, which is a widely accepted motor learning theory, based on distinct cognitive processes involved at different stages of skill execution (Fitts & Posner 1967). Initially, trainees are in a cognitive stage, during which they are attempting to understand what is to be performed. The next step is the associative stage, where progress slows as the trainee begins to modify movement strategies based on feedback. The last step is the autonomous stage where motor movements are performed automatically, requiring less attentional capacity. Performance becomes more fluid with increased experience (Reznick & MacRae 2006). Another descriptive step-wise model defining the development of technical skills is the Dreyfus and Dreyfus model, in which one progresses from novice to expert through five steps (Dreyfus 2004). See Table 1 for an overview of the theoretical models.

Characterization of expertise

Often, expertise has been equated with repeated, deliberate practice (Ericsson 2015). However, evidence indicates that this concept is flawed because it does not take into account individual differences (Kulasegaram et al. 2013; Macnamara et al. 2014). The number of completed procedures is not necessarily a comprehensive measure of expertise. In obstetrics, measures of initial skills have shown to explain more of the differences between individuals than procedural volume (Epstein et al. 2013). The development of expertise is a complex process and several factors have been proposed which may account for these individual differences, including working memory capacity. Further characterization of the individual differences that impact skill acquisition remains to be elucidated. Nevertheless, practice of technical skills is still a central part of achieving surgical proficiency and an extensive amount of evidence exists on efficient instructional strategies.

Efficient instructional strategies

Several instructional strategies have been shown to maximize the effect of training (Cook et al. 2013a, Stefanidis et al. 2014). On numerous occasions, it has been demonstrated – also for complex motor skills – that skill practice is most effective if it is structured as multiple training sessions, of short duration, spaced over time (distributed training) and with variable task practice (Nicholls et al. 2016). Feedback to the trainee also impacts the learning outcome significantly (McGaghie et al. 2010). Part-task training is also recommended for some complex tasks/procedures, depending on the level of interactive elements (Spruit & Band 2014). In order to teach complex technical skills, it is essential to take the current abilities of the trainee into consideration (cognitive load awareness; cf. Table 1; Spruit & Band 2014; Nicholls et al. 2016). One widely accepted approach is proficiency-based training, which involves a process of continued training until mastering a pre- and well-defined skill set (McGaghie et al. 2010).

Proficiency-based training

This method has shown to be efficient for learning technical skills and ensures that all trainees reach a minimum level of competency. This approach is gaining popularity in favour of time-based or repetition-based training where all trainees train for the same period of time or the same number of repetitions. This latter approach may lead to highly variable skill levels. Proficiency-based training requires: (1) explicit characterization of performance goals – how can proficiency be defined for this specific procedure? (2) assessment of performance – how can proficiency be measured? (3) establishment of benchmark criteria – when is proficiency reached?

Mastery learning is a rigorous approach to proficiency-based training and includes baseline testing, clearly defined learning objectives, establishment of a minimum passing benchmark, formative testing (continuous feedback), advancement to next educational unit when mastering the predefined benchmark level, and continued practice (Cook et al. 2013c, McGaghie et al. 2014). This approach is closely related to the deliberate practice model, as defined by Ericsson (2015).

Assessment of technical skills

It is necessary to assess performance in order to implement proficiency-based training, but assessment of performance is also critical for evaluating milestones and competencies. The latter is becoming an increasingly larger part of surgical practice, especially during residency (Oetting 2009, Accreditation Council for Graduate Medical Education 2014). Moreover, for research purposes, assessment is necessary to measure the effect of a training intervention.

Several rating tools for the assessment of technical skills exist. Overall, they can be grouped into three categories: (1) human rater tools, including global rating scales and procedure-specific checklists; (2) automated assessments, for example virtual-reality simulators and motion analysis; and (3) outcomes data, including surgical complication rates. The assessment tools differ in their level of objectivity and
how time-consuming they are to apply. Optimally, an assessment tool has several features: (1) validity and reliability evidence; (2) educational impact; (3) cost-effectiveness; and (4) widespread acceptance (Schuwirth & van der Vleuten 2013). In addition, some would add objectivity to this list (Gensheimer et al. 2013).

Automated assessment tools are objective assessment tools, which are not influenced by any pre-existing opinions about the trainee. Anonymized video recordings of trainee performances evaluated by masked raters also represent an objective assessment method.

All categories of assessment tools may be applied in both clinical and simulation-based settings. Simulation-based assessments often correlate positively with patient-related outcomes (Brydges et al. 2015). Therefore, in select cases, these assessments can replace or support those collected in the clinical setting (such as provider behaviours and patient outcomes). This possibility confers significant benefits because clinical assessments can be difficult to collect due to associated costs and logistical challenges (e.g. infrequent clinical events, nonstandardized settings, etc.). See Table 2 for a comparison between assessments performed in a simulated and clinical setting.

Validity evidence
Validity evidence – scientific evidence that the metric or test measures what it sets out to measure – is a requirement for useful, sound assessments of performance. The same applies for diagnostic tools in a clinical setting – we need to know whether a tool measures what we think it measures. For example, ear thermometry has been shown not to provide exact measurements of core body temperature (Stavem et al. 1997). Similarly, we have to investigate which measurements are useful for the assessment of technical skills.

When evaluating validity evidence for assessment outcomes, several different theoretical frameworks exist which outline a systematic approach. Messick’s validity framework is a widely respected approach, which has largely replaced the ‘classical’ framework, consisting of different types of validity (face, content, criterion, construct and concurrent). In Messick’s framework, validity evidence is collected from different sources, focusing on the intended use of the assessment. In most validity studies, measures with well-known characteristics are compared to new measures (in Messick’s framework: relations with other variables). For example, the correlation between surgical experience (traditional measure of surgical skills) and virtual-reality performance metrics (new measure) is investigated. However, to provide a robust argument for validity, it is often relevant to include multiple sources of validity evidence (Downing 2003); see Table 3 for an overview.

Unfortunately, validity evidence is often found to be limited, and in cases where the validity of performance outcomes have been investigated, most studies focus on extreme group comparisons (Cook et al. 2013d; Cook 2015). A typical example is a comparison between performance scores for a group of medical students and a group of experienced surgeons. A difference in performance score between the two groups is interpreted as evidence of validity based on the conception that more experience leads to superior technical skills (cf. the notion of deliberate practice) – however, the difference between the groups may reflect other differences than the measurement of interest (in this case, technical skills). For example, differences in anatomical knowledge may influence the results (confounding factor) and, therefore, lead to inaccurate conclusions. A more useful approach is the comparison of experienced surgeons to residents, who most often are similar to experienced surgeons except for their extent of surgical experience. The ability to discriminate between these two groups is a necessary, but not sufficient component of validity evidence (Cook 2015).

Another possible source of bias in validity arguments related to virtual-reality assessments is a familiarization effect. Specifically, it appears that individuals have different rates of familiarization to the virtual-reality interface, and therefore, it is crucial to include a warm-up period before collecting performance assessments. It is also important to remember that assessments are context-specific, and a validity argument may not necessarily be transferable from one environment to another; for example a human rater assessment tool, which has shown evidence of validity when used for video-recorded surgeries, may not provide meaningful measurements of proficiency when used for the assessment of technical skills in a wet-laboratory model, even if the same type of surgery is evaluated (Schuwirth & van der Vleuten 2011).

Reliability
Reliability is also a necessary property for assessment metrics and is part of the ‘internal structure’ in Messick’s framework, such that the measurements have to be consistent if repeated under similar conditions. In classical test theory, if trainees take an infinite number of tests measuring technical skills, the observed mean score $X$ is expected to be equal to the true score $T$, and the anticipated error $E$ has to be zero:

$$T = X + E$$

Table 1. Descriptive models for the acquisition of technical skills*.  

<table>
<thead>
<tr>
<th>Dreyfus stage</th>
<th>Description</th>
<th>Fitts-Posner stage</th>
<th>Performance characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice</td>
<td>Incomplete understanding, needs supervision to complete tasks</td>
<td>Cognition</td>
<td>Erratic, distinct steps</td>
</tr>
<tr>
<td>Advanced beginner</td>
<td>Working understanding, can complete simpler tasks without supervision</td>
<td>Integration</td>
<td>More fluid, fewer interruptions</td>
</tr>
<tr>
<td>Competent</td>
<td>Good working and background understanding, can work independently to an acceptable standard</td>
<td>Automation</td>
<td>Continuous, fluid, adaptive</td>
</tr>
<tr>
<td>Proficient</td>
<td>Deep understanding, can achieve a high standard routinely</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert</td>
<td>Authoritative understanding, deals with routine matters intuitively, achieves excellence with ease</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Adapted from Dreyfus (Dreyfus 2004) and Reznick et al. (Reznick & MacRae 2006).
The levels can be changed to estimate reliability under different testing conditions, called a decision study (D-study). Conceptually, reliability in generalizability theory is the trainee variance divided by the trainee variance plus the absolute or relative error variance. If the variance due to trainee differences is small and the error variances (facets) are large, the estimate of the generalizability will be a small number, and the score becomes less reliable.

When sources of validity and reliability evidence of the assessment metrics (e.g. performance test) have been evaluated, the next step is to define when proficiency has been reached when applying the assessment to training programmes.

Several analytical models exist to calculate an appropriate level for the intended purpose, but currently, there is no gold standard. One proposed method is the contrasting groups’ method, where the criterion is based on the intersection of test score distributions for the groups of novices and experienced surgeons (Downing & Haladyna, 2009).

Table 2. Simulator versus clinical setting for the training and assessment of technical skills.

<table>
<thead>
<tr>
<th>Assessment categories</th>
<th>Simulator environment</th>
<th>Clinical setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time requirement</td>
<td>Human rater tools/automated assessment</td>
<td>Human rater tools/automated assessment (motion analysis)/ outcomes data</td>
</tr>
<tr>
<td>Assessment setting</td>
<td>Less time required by educators</td>
<td>Time-intensive</td>
</tr>
<tr>
<td>Repetitions</td>
<td>Standardized</td>
<td>Dependent on patient cases</td>
</tr>
<tr>
<td>Variability</td>
<td>Repeated practice possible</td>
<td>Dependent on frequency of specific events</td>
</tr>
<tr>
<td>Part-task training</td>
<td>Variable task training possible incl. various difficulty levels</td>
<td>Possible but requires change of surgeon</td>
</tr>
<tr>
<td>Fidelity (level of authenticity)</td>
<td>Safe environment (no patients)</td>
<td>Patient safety concerns</td>
</tr>
</tbody>
</table>

Table 3. Sources of validity evidence*.

<table>
<thead>
<tr>
<th>Source</th>
<th>Definition</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content</td>
<td>Relevance of test content when compared to domain of interest</td>
<td>Ensuring representative assessment, for example by conducting review by experts</td>
</tr>
<tr>
<td>Response process</td>
<td>Evidence of data integrity (all sources of error associated with test administration are controlled or eliminated)</td>
<td>Controlled assessment environment, that is masked outcome assessors/raters</td>
</tr>
<tr>
<td>Internal structure</td>
<td>Reproducibility and generalizability of the test</td>
<td>Reliability of the assessment tool, that is intermodule reliability analysis using intraclass correlation coefficients</td>
</tr>
<tr>
<td>Relations with other variables</td>
<td>Correlation with external, independent measures</td>
<td>Assessment results related to previous surgical experience</td>
</tr>
<tr>
<td>Consequences</td>
<td>Consequences of test use</td>
<td>Considerations regarding passing rates including documentation of the method used to establish a pass/fail score</td>
</tr>
</tbody>
</table>

* Modified from Downing & Haladyna (2009).
Haladyna 2009). Another proposed method uses data from experienced surgeons, to ensure that trainees reach a level of automaticity, see Table 1 (Stefanidis et al. 2012). This does not mean that proficient trainees are considered to have reached the same level as the experienced surgeons, but rather it is interpreted as acquisition of an adequate level of competence on a particular training model. A third method – though not traditionally used for standard setting – involves analysis of novices’ performance curves, such that the plateau on their individual performance curve is identified and used as proficiency criterion. See Table 4 for an overview of standard setting methods. In most cases, identification of the most useful standard setting method depends on the specific aim of the proposed test and the particular target group for which the test is developed.

Outcome measures – efficacy studies

The next question to be answered is: what is a relevant training outcome? Is it relevant that a trainee can demonstrate improved performance on a simulator model – the same model on which they have trained repeatedly?

Miller’s framework for clinical assessment which extends from two cognitive levels (‘Knows how’) to two behavioural levels (‘Shows how’ and ‘Does’), mirrors on Kirkpatrick’s four levels of evaluation, see left triangle in Fig. 1 (Miller 1990). From this model, it is evident that changes in behaviour or results (i.e. patient outcomes) are the most valuable outcomes to demonstrate the potential effect of a training intervention. Cost-effectiveness studies (or transfer-effectiveness ratio) have been proposed as the highest level of evidence (Stefanidis et al. 2014).

Central concepts in efficacy studies include retention and transfer of skills. Transfer of acquired skills may be evaluated from one procedure to another or from one modality to another, for example from a virtual-reality environment to the operating room. The transfer from an in vitro environment to another setting may be defined as translational research, and the outcomes may be subdivided into the following categories: (T1) evaluation in simulation environments (i.e. on different tasks/procedures/models); (T2) improved patient care practices in a clinical setting; (T3) enhanced patient outcomes; and (T4) collateral effects, such as cost savings (McGaghie et al. 2014).

Transfer of skills from one modality to another (virtual-reality to OR)
The highest possible clinical impact of skills training is the documentation of a transfer effect from a simulation environment to improved patient care (provider behaviour) or improved patient outcomes, that is level T2 and T3 (Dawe et al. 2007). Specifically for virtual-reality interventions, only a few studies have investigated the training effect on patient outcome level (T3) in ophthalmology, and they show divergent results. Even though contributing with interesting and important findings, most existing studies are retrospective, and none of them report the use of standard instructional methods (Rogers et al. 2009; Belyea et al. 2011; Pokroy et al. 2013).

The transfer effect of virtual-reality training on provider behaviour in the operating room has been investigated in three trials (T2). One trial with a retrospective design finds a significant increase in number of errant capsulorhexis after implementing a capsulorhexis-intensive curriculum for residents (N = 38) on the EyeSi simulator (15.7% versus 5.0% in the postintervention cohort; p < 0.0001; McCannel et al. 2013). The two remaining studies are randomized controlled trials and conclude that virtual-reality training is equally effective to training on silicone eyes (N = 21) and patients (N = 8) when evaluating residents’ clinical performance (provider behaviour) for capsulorhexis and retinal photocoagulation performance, respectively (Peugnet et al. 1998; Daly et al. 2013). Peugnet et al. also found a significant decrease in training time, favouring the virtual-reality trained group over the patient-trained group.

Other studies have investigated the transfer of skills from one simulation model to another (T1). In particular, one study found that cataract surgical skills obtained in a virtual-reality environment are transferable to an animal model (porcine eyes). By contrast, two similarly designed studies found insignificant results and concluded that there was no transfer of skills for corneal suturing and posterior segment procedures (Feudner et al. 2009, Feldman et al. 2007, Jonas et al. 2003).

Transfer of skills from one procedure to another

Other factors are important when one consider the possible transfer effect from one type of procedure to another. According to Ericsson, skill development is domain- and task-specific, as supported by the specificity of practice hypothesis: skills acquired from practicing one procedure are not transferable to other procedures. This is explained by the principle of information processing specificity which states that the acquisition of skills in different procedures requires separate information processing (Grierson 2014). But how do you define the difference between task variability, which has been shown to have a positive effect on training outcomes, and tasks/procedures that are so dissimilar that no skill transfer exists? Studies investigating the importance of practice specificity in various domains suggest that the two tasks must be almost identical for transfer to occur (Kirkman 2013). This is demonstrated in a study by Selvan der et al., where no skill transfer is evident between two modules on the EyeSi simulator (specifically, the capsulorhexis and navigation module). Skill transfer from one surgical procedure to another has never been investigated in ophthalmic surgery.

Surgicalsimulators in ophthalmology

Simulation models can be categorized into four different groups: animal, cadaver, inanimate, and virtual-reality models. Traditionally, animal models have been used for the training of technical skills in ophthalmology, as they are relatively accessible and often represent a moderate to high fidelity (authenticity) compared to the human model. However, disadvantages include infection risk (including MRSA) and dissimilarities in biomechanical properties. Inanimate models encompass a wide variety of possibilities, including capsulorhexis training on red globe grapes, aluminium foil and mannequin heads, such as the Kitaro model. The advantage of using inanimate models is that they are readily available, relatively inexpensive and not associated with infection risk. However, the similarity
with the clinical setting may only be moderate, and training is often limited only to specific steps of a procedure.

**Virtual-reality simulators**

In ophthalmology, several virtual-reality systems have been developed but the most commonly used virtual-reality simulator is the EyeSi surgical simulator. This consists of a mannequin head, instruments, foot pedals and a virtual-reality interface, which is seen through the operating microscope. Both a cataract and a vitreoretinal interface are available. The cataract interface consists of all the major steps of the cataract procedure (procedural modules, enabling part-task training), in addition to abstract modules, which train basic skills. All modules can be used on different difficulty levels. However, this allows only for the training of intraocular tasks (i.e. incisions are not included). For an overview of the modules on the cataract interface, see Table 5.

The automated assessment provided by the simulator consists of 21–33 different outcomes categorized into five main outcomes: target achievement, efficiency, instrument utilization, tissue damage, and microscope usage. Parts of the scoring system are based on motion-tracking technology, which involves the measuring of millimetre movement of instrument tips and the number of instrument closings (i.e. how often a forceps is closed). Efficiency is a time-based measure, which is based on the total time that instruments are inserted in the eye. Not all modules include an efficiency measure, and in particular, the capsulorrhexis module is not time based, as precision is deemed more relevant than efficiency during this specific task.

**Hypothesis**

I The failure to use evidence-based methods in ophthalmology simulation research has led to inadequate knowledge in the field.

II Automated assessment of technical skills (including motion-tracking technology) with select modules on the EyeSi virtual-reality simulator and during real-life cataract surgery is associated with validity evidence.

III Proficiency-based virtual-reality training has a significant impact on novice surgeons’ performance level in the operating room.

IV The acquisition of complex technical skills is task-specific so no transfer of skills will occur from one intraocular procedure to another (specifically from cataract to vitreoretinal surgery).

**Research Objectives**

The overall aim of this thesis is to investigate the assessment and transfer effects of technical skills in intraocular surgery, including:

- Reviewing the evidence behind simulation-based surgical training of ophthalmologists to determine (1) the validity of simulation-based assessment metrics; and (2) the training effect on all outcomes, including the ability to transfer skills to the operating room (STUDY I)
  - **Methodology**: Systematic review

- Determining validity evidence for automated assessment of cataract surgical skills on an intraocular virtual-reality simulator (STUDY II/III)
  - **Methodology**: Cohort/cross-sectional studies

- Investigating transfer of cataract surgical skills from a virtual-reality environment to the operating room (STUDY IV)
  - **Methodology**: Interventional (before and after) study – multicentre and rater-masked

- Investigating transfer of skills between two intraocular surgical procedures; specifically from cataract to vitreoretinal surgery (STUDY V)
  - **Methodology**: Randomized controlled trial

**Summary of Studies**

**STUDY I**: Update on simulation-based surgical training and assessment in ophthalmology – a systematic review

**Background:**

Research on the acquisition and assessment of surgical skills in
ophthalmology using simulation models is relatively new. A systematic and evidence-based approach to the development and evaluation of surgical skills may ultimately result in improved patient safety.

**Objective:**
To review the evidence behind simulation-based training and assessment of surgical skills in ophthalmology.

**Methods:**
Eligibility criteria: We conducted a systematic review of all original trials investigating simulation-based surgical training or assessment methods related to the development of technical skills in ophthalmology.

Information sources: We searched five databases for eligible trials and included trials were divided into three key categories according to the purpose of the trial: (1) Description of a training or assessment model; (2) Validity investigation; (3) Efficacy measurement of skills training (including skill transfer to the operating room).

Summary measures: The evidence level of all included trials was primarily evaluated based on the Kirkpatrick model (see Fig. 1). A 'risk of bias' analysis, based on the Cochrane Handbook for Systematic Reviews of Interventions (version 5.1.0), was also conducted.

**Results:**
We screened 1368 reports for eligibility and included 118 trials in our final analysis. Cataract surgery was the most frequent procedure simulated (53%), followed by a group of trials with non-specific or mixed procedures (18%), and finally vitreoretinal surgery (13%). Most participants were residents in ophthalmology (42%), followed by medical students (24%). Most of the included trials were descriptive trials (61%) and described 60 different training models without further investigations.

Data were not suitable for statistical pooling and were included in a qualitative analysis due to heterogeneity in methodology and outcomes.

Overall, 23 trials investigated validity evidence of various assessment tools using nine different simulation models: three virtual-reality simulators, three animal models and three inanimate models. The assessment method categories included both automated assessments provided by virtual-reality simulators, motion-tracking systems, human rater evaluations, and participant questionnaires (assessing participant satisfaction). Most validity trials (74%) used extreme groups comparison and compared performance metrics for medical students or residents with experienced surgeons. These studies also addressed only one or two of the five sources of validity in Messick’s framework (see Table 3). The EyeSi simulator was investigated in 11 trials. All of these trials included only single modules without explicitly addressing the content of the assessment programme. None of the trials examined the correlation between simulation-based assessments and provider behaviour or patient-related outcomes (Kirkpatrick level 3 and 4). Seventeen trials examined the efficacy of simulation-based training. Twelve efficacy trials (71%) used the EyeSi simulator (cataract interface, N = 11). Most of the efficacy trials were designed as cohort trials (65%) and represented level 2 in the Kirkpatrick model (65%) by investigating whether repeated training on one model leads to improved performance on the same model. None of the trials reported or discussed validity evidence of the outcome measures used.

**Table 5. Overview of modules on EyeSi cataract interface**

<table>
<thead>
<tr>
<th>No.</th>
<th>Name</th>
<th>Levels</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Navigation</td>
<td>3</td>
<td>Abstract</td>
<td>Moving tip of instrument into spheres in anterior chamber</td>
</tr>
<tr>
<td>2</td>
<td>Navigation – intracapsular</td>
<td>3</td>
<td>Abstract</td>
<td>Moving tip of instrument into spheres placed in an empty capsule</td>
</tr>
<tr>
<td>3</td>
<td>Antitremor</td>
<td>7</td>
<td>Abstract</td>
<td>Follow path as precise as possible on anterior capsule</td>
</tr>
<tr>
<td>4</td>
<td>Antitremor – intracapsular</td>
<td>5</td>
<td>Abstract</td>
<td>Follow path as precise as possible on posterior capsule</td>
</tr>
<tr>
<td>5</td>
<td>Forceps</td>
<td>4</td>
<td>Abstract</td>
<td>Manipulation of objects in anterior chamber with forceps</td>
</tr>
<tr>
<td>6</td>
<td>Binamuel training</td>
<td>5</td>
<td>Abstract</td>
<td>Use two instruments to touch both ends of objects in anterior chamber</td>
</tr>
<tr>
<td>7</td>
<td>Cracking and chopping</td>
<td>8</td>
<td>Abstract</td>
<td>Chopping exercise</td>
</tr>
<tr>
<td>8</td>
<td>Phaco training</td>
<td>3</td>
<td>Abstract</td>
<td>Use phaco-machine to remove lens part</td>
</tr>
<tr>
<td>9</td>
<td>Capsulorhexis</td>
<td>3†</td>
<td>Procedural</td>
<td>Perform capsulorhexis with forceps after injecting viscoelastica</td>
</tr>
<tr>
<td>10</td>
<td>Hydromanuevre</td>
<td>8</td>
<td>Procedural</td>
<td>Perform hydrodissection and hydrodelineation</td>
</tr>
<tr>
<td>11</td>
<td>Phaco divide and conquer</td>
<td>8</td>
<td>Procedural</td>
<td>Perform sculpting, split lens into four quadrants and remove content</td>
</tr>
<tr>
<td>12</td>
<td>Irrigation and aspiration</td>
<td>5</td>
<td>Procedural</td>
<td>Remove cortex using I/A setting on phaco-machine</td>
</tr>
<tr>
<td>13</td>
<td>IOL insertion</td>
<td>4</td>
<td>Procedural</td>
<td>Inject viscoelastica, place intraocular lens and remove viscoelastica</td>
</tr>
</tbody>
</table>

* Version 2.8.10.
† Weak zonula. No initial tear.

**Transfer of skills to the operating room/clinical setting:** Three trials investigated transfer of surgical skills to the operating room or clinical setting, representing the second highest level in the Kirkpatrick model. In these trials, virtual-reality training was found to be equally effective or superior to traditional training (Peugnet et al. 1998; McCannel et al. 2013) and equally effective when compared to training on an inanimate model (Daly et al. 2013). However, all trials were associated with a significant risk for bias and evaluated retinal photocoagulation and the cataract surgical task capsulorhexis.

**Patient outcomes:** Four trials evaluated the effect of a simulation-based training intervention on patient outcomes and achieved the highest level of evidence in the Kirkpatrick model. All of these trials investigated the effect of new surgical curricula – including training on the EyeSi virtual-reality simulator – on procedural time and complication rates in an operating room setting. Three of them were based on a retrospective design and, therefore, had a relatively high risk of bias and confounding. The final study was a case series of three residents. Overall these trials had some conflicting results.
Two of them demonstrated a significant decrease in cataract surgical complication rates (posterior capsule tear and/or vitreous loss) for residents ($N = NA$) but this finding was not confirmed by the two other trials ($N = 62$). There was some evidence for reduction in procedural time after the implementation of new surgical curricula.

**Conclusions:**
- Numerous simulation models have been introduced but have not been investigated further (neither as assessment tools or training models).
- There is a lack of evidence supporting the use of simulation-based assessments, and in particular, knowledge regarding the correlation with provider behaviour or patient-related outcomes is lacking (validity studies).
- Validity of the EyeSi automated assessment metrics has not been well established.
- Efficacy trials show a tendency towards improving surgical time and complication rates after training on the EyeSi simulator; however, data are inconsistent so that no final conclusions can be made.
- Future studies should focus on improving the quality of their design, including establishing validity evidence for outcome metrics, applying evidence-based instructional methods, and minimizing risk of bias and confounding.

**STUDY II: Simulation-based certification for cataract surgery**

**Background:**
Simulation-based training may make it possible to minimize the need to learn by performing procedures on patients. But how do we decide how much practice is enough?

**Objective:**
Determining validity evidence for automated assessments of cataract surgical skills using a virtual-reality simulator and setting a benchmark criterion for proficiency intended for surgical novices.

**Methods:**
**Study design and setting:** The study was carried out as a cohort study at the Copenhagen Academy for Medical Education and Simulation, Denmark from November 2013 to April 2014.

**Participants:** Three groups of participants were included in the study: (1) novices (employed at an ophthalmology department without any cataract surgery experience), (2) experienced cataract surgeons ($>4000$ cataract procedures), and (3) experienced vitreoretinal surgeons ($>200$ vitreoretinal procedures and limited cataract surgery experience).

**Variables:** We constructed a test reflecting learning objectives for residents in a cataract surgery curriculum. This test consisted of all cataract surgical modules on the EyeSi simulator (VRmagic, version 2.8.10) with the exception of one alternative procedural approach (chopping), which was not deemed relevant for a cataract surgical trainee.

**Outcome measures:** All participants completed two test sessions after a warm-up period of 10 min. Target achievement, efficiency, instrument utilization and tissue treatment were automatically assessed by the simulator and resulted in a module score with a maximum 100 points.

**Results:**
A total of 42 participants completed the study: 26 novices, 11 cataract surgeons, and five vitreoretinal surgeons. All participants improved significantly from their first to second attempts, indicating a familiarization effect. For this reason, we decided to use data only from the second attempt. Seven of the 13 modules demonstrated statistically significant discriminative ability between novices and experienced cataract surgeons (one-tailed independent samples $t$-tests) and good reliability between modules (ICC 0.76, $p < 0.001$). These seven modules were included in a final performance test.

**Mean performance scores (SD) for the final test were 333 (96) points, 462 (68) points and 497 (52) points for novices, vitreoretinal surgeons, and experienced cataract surgeons, respectively, with a statistically significant difference between novices and experienced cataract surgeons ($p < 0.001$), and between novices and vitreoretinal surgeons ($p = 0.006$). There was no statistically significant difference in mean performance test score between vitreoretinal surgeons and experienced cataract surgeons ($p = 0.32$).

A proficiency level (pass/fail score) was determined at 422 points using the contrasting groups’ method. **Post hoc** consequence analysis showed that the final performance test was effective in distinguishing between novices and experienced cataract surgeons, as 21 (81%) novices did not pass the proficiency test versus one (9%) of the experienced cataract surgeons.

**Conclusions:**
- We have established a performance test, consisting of seven modules on the EyeSi simulator, which possesses evidence of validity (4 of 5 sources investigated).
- A proficiency level of 422 points was determined and may be used for future implementation of proficiency-based training for surgical novices.

**STUDY III: High correlation between performance on a virtual-reality simulator and real-life cataract surgery**

**Background:**
The EyeSi virtual-reality simulator is an accepted tool for the training and assessment of surgical skills – primarily for cataract surgery – but there is a lack of evidence supporting the direct translation between the automated assessment provided by the simulator and real-life surgical performance.

**Objective:**
To investigate the correlation between virtual-reality performance and real-life cataract surgical performance using a simple model, which compares motion-tracking metrics from real-life cataract surgery to performance metrics from the EyeSi simulator.

**Methods:**
**Study design and setting:** A cross-sectional study was conducted at ophthalmology departments and private ophthalmology clinics in Denmark and at the Copenhagen Academy for Medical Education and Simulation, Denmark. Data were collected from April 2014 to March 2015.

**Participants:** Cataract surgeons – operating independently, but with different levels of surgical experience – were included in the study.

**Variables:** We collected videos of three uncomplicated cataract surgeries consecutively performed by each surgeon prior to their simulation session. Uncomplicated cataract surgeries were defined as: (1) surgery performed under local anaesthesia, (2) patient age $> 60$ years, and (3) visual acuity $> 1/60$ pre-operatively (measured using a...
standard Snellen chart at 1 m distance). The videos were analysed using motion-tracking parameters, which have previously demonstrated evidence of validity (Smith et al. 2013). During the 2 weeks after video collection, all participants completed an evidence-based proficiency test of cataract surgical skills (see STUDY II) on the EyeSi simulator (VRmagic, version 2.8.10).

Outcome measures: The primary outcome measures were (1) summative virtual-reality performance test score (collected after two warm-up sessions, corresponding to approximately 1 hr) and (2) motion-tracking score from real-life videos, calculated by average number of movements based on three full-length live cataract surgeries.

Results:
In total, eleven cataract surgeons were included in the final data analysis; two surgeons were excluded due to technical issues with video recordings. The number of cataract surgeries previously performed by each surgeon varied from two to 24,220 (mean 3656) and videos were collected from six ophthalmology departments and private ophthalmology clinics in Denmark.

Motion-tracking scores varied from 34,770 to 652,521 (a lower motion-tracking score reflecting superior performance), while the virtual-reality performance test scores ranged from 485 to 661 points out of a maximum of 700 points. We found a significant association between virtual-reality performance test scores and motion-tracking scores (p = 0.017, adjusted R² = 0.43) using a linear regression model. A ten-unit improvement in the virtual-reality performance test score was estimated to correspond to a decrease of ~23 220 in the motion-tracking score (95% confidence interval: -41 240, -5210).

Conclusions:

- Performance on the EyeSi simulator is highly correlated with real-life surgical performance and may supplement workplace-based/clinical assessments.
- Motion-tracking metrics are associated with high variance between individuals, which is most significant for surgeons with an intermediate level of experience. This is probably explained by an experimental movement pattern (i.e. exploring different surgical techniques).
- Therefore, multiple data sources are still recommended when evaluating surgical skills.

STUDY IV: Operating room performance improves after proficiency-based virtual-reality cataract training

Background:
Ideally, surgical skills acquired in a simulation-based setting are transferable to the operating room. However, it remains unclear whether virtual-reality training leads to superior surgical performance and this has never been investigated prospectively for the entire cataract surgical procedure. Also, we do not know if a potential training effect exists only for novices or if more experienced surgeons benefit from virtual-reality training.

Objective:
To investigate the effect of proficiency-based, virtual-reality training on cataract surgical skills in the operating room for surgeons with different levels of experience.

Methods:
Study design and setting: A multicentre, national, prospective clinical trial (cohort study) with masking of findings of the participant’s next surgery performed, (coemulsification surgeries, immediately post hoc to the lowest score (‘inadequately performed’).

Results:
In total, 18 surgeons were included in the final data analysis; one experienced surgeon was excluded due to missing data. Four surgeries were excluded and replaced by video recordings of the participant’s next surgery due to technical issues (N = 2) or a complicated surgical procedure (N = 2).

The generalizability coefficients for the performance assessments were 0.92 and 0.86 for pre- and post-training, respectively. The reliability analysis showed that only a small proportion of the total variance was due to differences in case difficulty, raters, or procedure-to-procedure variance.

Novice, intermediate, experienced, and expert surgeons had pretraining mean performance scores of 15.33, 25.81, 42.97, and 48.26 points (out of a total of 52 points), respectively. Both novices and intermediates showed statistically significant improvements in performance, by 5.0 and 9.8 points corresponding to an increase of 32% and 38% from baseline values, respectively (p = 0.008 and p = 0.018, mixed-effects regression analysis). There were no significant improvements for experienced and expert groups.

Conclusions:

- A proficiency-based virtual-reality training programme improves surgical performance by 32% in novice cataract surgeons.
• Intermediate level surgeons (who have performed <75 independent operations) also benefit significantly from virtual-reality training (38% improvement).
• Improvement in technical skills resulting from proficiency-based training seems to be transferable from a simulated setting to the operating room.

STUDY V: Is there interprocedural transfer of skills in intraocular surgery? a randomized controlled trial

Background:
One may expect that skills learned for one procedure may be transferable to another related procedure. However, the required level of similarity between procedures to detect a procedure-to-procedure transfer effect is unknown.

Objective:
To investigate how proficiency in virtual-reality cataract surgery affects learning curves for novices in vitreoretinal surgery.

Methods:
Study design and setting: The study was carried out as a randomized controlled trial (balanced randomization 1:1). Data were collected at the Copenhagen Academy for Medical Education and Simulation, Denmark, from October 2014 to June 2015.

Participants: Ophthalmology residents with no intraocular surgical experience were recruited from the Department of Ophthalmology, Rigshospitalet – Glostrup, Denmark.

Intervention and control: Residents were randomized to two different groups: (1) participants who passed an evidence-based cataract surgical performance test on the EyeSi simulator during two consecutive sessions (see STUDY II) or (2) participants who received no intervention.

Outcome measures: Participants in both study arms repeated a vitreoretinal performance test until reaching their maximum performance level (plateau level), defined as two consecutive sessions with a score <5% higher than the highest score previously obtained. The vitreoretinal performance test consisted of all 11 vitreoretinal modules on the EyeSi surgical simulator (VRmagic, version 2.8.10). The primary outcomes were total test score from the first attempt (after approximately 1 hr warm-up), total time to reach maximum performance level, and maximum score. All outcomes were based on the automated assessments provided by the EyeSi virtual-reality simulator.

Results:
We included twelve ophthalmology residents; six participants were allocated to intensive cataract surgical training (cataract trainees) and six participants received no training (novices). Additionally, we included three experienced vitreoretinal surgeons to establish validity evidence for the outcome measure.

Validity evidence for the outcome measure: The automated assessment metrics showed statistically significant discriminatory ability between experienced vitreoretinal surgeons and novices (p = 0.003) and cataract trainees (p = 0.006) for the first attempt. The intermodule correlation coefficient was 0.73 (Cronbach’s alpha).

Procedure-to-procedure skill transfer: The group of cataract trainees did not outperform the group of novices when comparing starting score (mean ± SD 381 ± 129 points versus 455 ± 82 points, p = 0.26); time to reach maximum performance level (10.7 ± 3.0 hr versus 8.7 ± 2.8 hr, p = 0.27); or maximum score (785 ± 162 points versus 805 ± 73 points, p = 0.79). The 95% confidence interval for the mean difference was –196 to 48 points with respect to the starting score; –1.3 to 5.3 hr for time to reach maximum performance level; and –162 to 122 for points in maximum score.

Proficiency level: Post hoc analysis showed no difference in mean scores for the last three repetitions for each group: F(2;10) = 3.3, p = 0.080; F(2;10) = 3.8, p = 0.058; and F(2;4) = 0.045, p = 0.96 for novices, cataract trainees and experienced surgeons, respectively (repeated measures ANOVA per group). Similar analysis demonstrated no significant difference in mean maximum scores between groups (F(2;12) = 0.102, p = 0.90).

Conclusions:
• All participants, independent of group, were able to reach the same maximum performance level on the simulator, indicating a ceiling effect for the automated assessment of technical skills on the EyeSi virtual-reality simulator.
• No significant transfer effect of surgical skills was found from cataract surgery to vitreoretinal surgery in a virtual-reality environment; our findings suggest a negative transfer effect.
• Significant transfer of surgical skills should not be anticipated when planning training programmes in intraocular surgery.

Discussion
Efforts to enhance the quality of surgical care should involve research that addresses the ‘surgeon factor’ – namely, the specific and varied needs of healthcare professionals in their training and development of clinical skills, including technical skills. Clinical performance outcomes may improve significantly with targeted training interventions and may be considered equally as important as other clinical studies, which assess the effectiveness of new treatments. This area is developing rapidly, as is evident from the recently published CONSORT and STROBE guidelines which promote the need for high-quality studies in simulation research (Cheng et al. 2016). Our systematic review reveals that there is room for improvement and, in particular, that validity evidence for performance metrics in ophthalmology is lacking. None of the trials included in our systematic review reported or discussed the validity of the outcome measures used. Moreover, it is clear that the use of evidence-based training and assessment methods has not received significant attention in ophthalmology training in Europe. The majority of educators – responding to a cataract surgical training survey – reported that they still use the apprenticeship model in cataract surgical training, as opposed to the proficiency-based model (Muttuvelu & Anderssen 2016). Thus, there seems to be a tendency to perceive surgical training as an expert domain, rather than as an evidence-based process. This may lead to costly, ineffective surgical training programmes, or could even present a threat to patient safety.

In this thesis, assessment methods (STUDY II–III), as well as the acquisition and transfer (STUDY IV–V) of technical skills in intraocular surgery have been investigated. Based on our study findings, we conclude that the automated assessment based on the EyeSi virtual-reality simulator metrics is useful in evaluating progress in cataract surgical skills, and can supplement, or even replace, workplace-
based/clinical assessments for this specific procedure. Outcome measures with evidence of validity are important when implementing proficiency-based surgical training, and in STUDY IV, we found that proficiency-based virtual-reality training had a significant impact on performance measures (provider behaviour) in the operating room for both novices and intermediates. While transfer of skills from a virtual-reality setting to the operating room seems to occur, skill transfer from cataract surgery to vitreoretinal surgery has not been established.

All of the studies share several strengths with respect to their design: the inclusion of relevant groups of participants (not medical students), as well as the use of evidence-based and contemporary methodology, including proficiency-based training for the acquisition of technical skills and Mesick’s framework for collecting validity evidence for performance metrics (Gallagher et al. 2005; Korndorffer et al. 2010).

The most important limitation of our study design is small sample sizes due to a limited number of cataract and vitreoretinal surgeons in Denmark. Potential possibilities to address this limitation could have been the inclusion of other microsurgeons or ophthalmologists in other countries. International data collection was planned for one of the studies, but differences in equipment and healthcare systems (including training of specialists) between the countries presented an insurmountable barrier.

Another important limitation is that in all the studies we only included technical skills as an outcome measure. In reality, a wider array of skills is fundamental to safe practice, including decision-making, communication, and leadership.

**Automated assessment of technical skills**

The automated assessment provided by the EyeSi virtual-reality simulator – comprising motion-tracking metrics, time and error measurements (tissue damage) – demonstrates evidence of validity and reliability for cataract surgical skills for all levels of experience when using select modules. Thus, the presented performance test may be implemented in cataract surgical training programmes to support the progress of novices towards proficient surgeons. At present, the number of completed cases is often used as a proxy measure of proficiency, in addition to nonstandardized evaluations made by senior colleagues (Muttuvelu & Andersen 2016). Both of these assessment methods introduce a considerable amount of bias. Traditionally, workplace-based assessments have been presented as the best way to assess professional competence, including technical skills (Govaerts & van der Vleuten 2013). Yet, workplace-based assessments are often cumbersome to collect both in regards to time and effort. Interestingly, Brydges et al. (2015) conclude from a meta-analysis that tools with established validity evidence may replace workplace-based assessments in the evaluation of select procedural skills. The findings of STUDY II and III in this thesis support the use of the presented virtual-reality performance test as a replacement for workplace-based assessments in cataract surgery.

STUDY II and III have some notable strengths in their design, related to the bi-faceted investigation of the construct ‘relations to other variables’. We investigated the EyeSi automated assessment metrics both in relation to experience level (STUDY II) – which is considered an indispensable but not sufficient part of the validity argument – and compared the EyeSi metrics to motion-tracking parameters from real-life videos (STUDY III). To our knowledge, this is the first study to assess the correlation between simulation-based assessments and provider behaviour in ophthalmology (Kirkpatrick level 3). In STUDY III, we chose not to use experience as a proxy measure for proficiency, and we are aware that adding experience to the model would probably have resulted in a higher value for $R^2$. However, this was not our intent, as we wanted to investigate the simple relationship between clinical motion-tracking parameters and the automated assessment provided by the EyeSi simulator.

The findings of these studies revealed that motion-tracking metrics – as used in STUDY III – might be useful as an objective measure of surgical proficiency in intraocular surgery. These metrics can differentiate between novices and experienced surgeons, but are suboptimal for the assessment of intermediate surgeons. This phenomenon may be explained by ‘experimental practice’, such that the surgeons with an intermediate level of experience tend to explore movement patterns as part of their learning process, and therefore, they may obtain lower performance scores than novices. Similar findings have been documented by Rojas et al. (2011) for knot-tying procedures. These findings support previous evidence indicating that end-product assessments are more predictive of overall competence than assessments focussing on the process (Schuwirth & van der Vleuten 2011).

The discussion points specifically relevant to these studies include target group, study context and specificity of the assessed skills. First of all, we set out to develop a test of cataract surgical skills for surgeons at the beginning of the learning curve (STUDY II), but the results from STUDY III suggest that the test may be utilized for surgeons at all levels of experience. Nevertheless, it is important to remember that most assessment tools show a ceiling effect for superior performance, thus often it is more challenging to discriminate between intermediates and expert surgeons than between novices and experienced surgeons. This pattern is also evident from our motion-tracking data. Future work should focus on the development and investigation of new benchmark criteria, which corresponds to the experience level of the defined target group. Overall, it is critical to give careful considerations to the defined target group, as each group has unique concerns and what may seem like an optimal assessment tool for novices may not be applicable for experienced surgeons. Moreover, one assessment tool should never stand alone as a measure of surgical proficiency. The EyeSi simulator provides useful knowledge regarding cataract surgical skill, but several assessment tools are recommended – especially for high-stakes assessments such as board exams (Boursicot et al. 2011; Schuwirth & van der Vleuten 2011).

Context is also important. We have to remember that evidence of validity and reliability is only applicable in the original context, that is corresponding to the circumstances under which the evidence was collected (Schuwirth & van der Vleuten 2011). An example is the measurement of blood pressure in different settings. When measured in
the clinic, a ‘white coat syndrome’ may lead to increased blood pressure compared to at-home measurements. Due to constant developments in virtual-reality software – in addition to possible changes in the environment and target groups – validity of performance metrics are never static and must never be interpreted as such.

Finally, a designated warm-up period is crucial for sound assessments of performance, and our findings from STUDY II suggest that approximately 1 hr of warm-up is essential for reliable assessments of performance on the EyeSi simulator. Failure to provide a warm-up period may result in an underestimation of expert performance, as differences in the technical aspects of a procedure in a simulation-based setting (as compared to the clinical setting) may affect performance initially. Other studies have reached similar conclusions (Schijven & Jaki-mowicz 2004); however, omitting warm-up periods remains a recurring issue in simulation research.

Regarding the specificity of the assessed skills, the findings of STUDY II suggest that the automated assessment is not specific for cataract surgical skills, but rather assesses intraocular surgical skills in general. The group of vitreoretinal surgeons did not perform significantly differently from the group of cataract surgeons. However, this finding may reflect a type II error due to a small sample size, as only three vitreoretinal surgeons were included in the study. The lack of specificity in the performance assessments may also be affected by a ceiling effect – slight differences in high levels of performance may not be distinguishable, and thus, a possible difference in performance between cataract and vitreoretinal surgeons may not be detected. Overall, if the lack of specificity is true, the differences found between novices and cataract surgeons (STUDY II), as well as between different levels of cataract surgical expertise (STUDY III), may be solely due to differences in exposure to microsurgery and not specifically cataract surgery. This leads us to the concept of skill transfer between procedures.

**Specificity of training interventions**

Given the specificity of practice theory, one should not expect any transfer of skill between different types of intraocular surgery because they involve different sequences of movements. This theory is supported by the findings of STUDY V, where cataract surgical training-to-proficiency does not influence the acquisition of vitreoretinal surgical skills for novices in a virtual-reality environment. Based on these findings, the result of STUDY II – showing no significant difference in performance assessments between experienced cataract surgeons and vitreoretinal surgeons – may be interpreted as a type II error compounded by a ceiling effect for high levels of performance.

From the results in STUDY V it actually seems that negative transfer occurs. Negative transfer refers to the finding that the applied skill training negatively affects end-point performance. In STUDY V, the cataract trainees spent on average two hours longer on the vitreoretinal interface before reaching their maximum performance level compared to the novices (no prior training). Thus, complex procedural skill development in intraocular surgery seems to be domain specific. The differences in surgical instruments and anatomical structures may explain the finding of negative transfer. However, even with the same instruments and within the same anatomical structures in the eye, Sel-vander & Asman (2012) found no evidence of skill transfer from an abstract module to a procedural task on the EyeSi simulator. Procedure-to-procedure transfer has not been previously investigated for entire ophthalmic procedures, and notably, in laparoscopic surgery, results are inconsistent (Bjerrum et al. 2015). Basic skills training using instruments similar to real life (black-box training) has shown to prepare trainees for thoracoscopic lobectomy more effectively than advanced high-fidelity simulation training in nephrectomy (Jensen et al. 2014). Overall, evidence suggests that for complex procedural skill development, it is difficult to support skill transfer between different types of procedures (Kirkman 2013). The challenge is to define various elements that seem important for skill transfer. There is a significant difference between the process of procedure-to-procedure transfer and transfer between different environments (or training modalities), and our understanding of these differences may benefit from efforts to distinguish between the functional features, such as sensory and motor information processing, and structural features (learning environment/setting) of these tasks (Grierson 2014). From these research findings, we may conclude that instrument and anatomical similarity, as well as movement patterns, are important factors for skill transfer, corresponding to functional features of the tasks. Structural alignment seems less important in this regard (Norman et al. 2012).

Limitations specific to STUDY V include the target group and study setting. This study was conducted in a virtual-reality setting and included only surgical novices. Skill transfer from cataract to vitreoretinal surgery may be evident for more experienced surgeons. However, previous studies suggest that negative transfer may be even more pronounced for experienced surgeons as their associative memory is stronger compared to novices (Starkes et al. 1993). At the same time, positive procedure-to-procedure transfer may exist for more experienced surgeons as their additional cognitive resources make it possible to have skill transfer in situations where novices will experience cognitive overload (Grierson 2014). Importantly, in our study, the procedure-to-procedure transfer of skills was only investigated in one direction, and the opposite direction of skill transfer, for example from vitreoretinal to cataract surgery, may not be excluded.

Other conditions apply when it comes to the same skills set, but transfer between different environments. Evidence suggests that transfer occurs from one modality to another, that is skills developed on a virtual-reality simulator have been shown to be transferable to clinical performance or other simulation models requiring the same set of skills. This transfer has been confirmed based on improvements in provider behaviour. In ophthalmology, this has been established for capsulorhexis performance in a retrospective study (McCannel et al. 2013). There are no additional studies on this subject in ophthalmic surgery, but similar findings have been found for laparoscopic surgical skills (Aggarwal et al. 2007).

Additionally, we find that skills acquired from proficiency-based
training of abstract and procedural cataract surgical tasks in a virtual-reality environment are transferable to overall cataract surgical performance in the operating room. This corresponds to a translational research level of T2. These findings add to the body of evidence, suggesting that learning environment/setting, that is structural alignment, is less important than content for the transfer of complex procedural skills to occur (Hamstra et al. 2014).

The training of technical skills in a virtual-reality environment may be criticized for creating an oversimplification of a complex reality. On the other hand, the oversimplification provides the opportunity to concentrate on select skills – as explained by the cognitive load theory – possibly providing the opportunity to develop automaticity as explained by Fitts&Posner’s motor skill theory (Reiner & Gelfeld 2014). By achieving automaticity, it may become possible to focus attention on other important tasks when entering the operating room, such as communicating with the patient and surgical team, or dealing with the unexpected (Kurahashi et al. 2011; Stefanidis et al. 2012).

Intermediate level surgeons and virtual-reality training

Simulation-based training has obvious implications for novice surgeons and this may be the reason why there has been a tendency to focus on novices in simulation research. In STUDY IV, we wanted to investigate the ‘sensitivity to change’ among surgeons of different experience levels. Does transfer of skills from a virtual-reality setting occur only for novice surgeons, or do more experienced surgeons also benefit from virtual-reality training?

Previous studies have investigated the effect of warm-up sessions on virtual-reality simulators immediately prior to surgery and found that warm-up is associated with a significant improvement in performance, also for experienced surgeons (Moldovanu et al. 2011; Lee et al. 2012). Similarly, in our study, we found a significant effect of training also for surgeons with an intermediate level of experience. Based on these findings, there is a wide range of applications for virtual-reality training.

The defined benchmark criteria used for proficiency-based training seems crucial when utilizing training for more experienced surgeons. In STUDY IV, we used a higher pass/fail criterion than calculated in STUDY II to support ‘overlearning’, which seems to be a major factor for skill retention. This phenomenon is probably explained by the fact that participants reach a level of automaticity (Arthur et al. 1998; Stefanidis et al. 2006, 2012).

STUDY IV is a multicentre study, which increases the generalizability of the findings. An important limitation of the study is the lack of a control group. It would have been desirable to conduct a randomized controlled trial, but this was not feasible due to the limited number of cataract surgeons in Denmark. We included three masked raters and six different time-points for clinical assessments (three before and three after virtual-reality training), making advanced statistical analysis feasible (generalizability theory and mixed-effects regression analysis) and thus supporting the before-and-after study design.

Guidelines for simulation-based training in intraocular surgery

It has been suggested that simulation-based training – under some conditions – is even more effective than traditional training (Peugnet et al. 1998; McCannel et al. 2013; Konge et al. 2015). This may be explained by more efficient hands-on training time in a simulator environment as compared to clinical practice and makes simulation-based training even more attractive. There is growing evidence supporting the implementation of simulation-based training in ophthalmology, especially virtual-reality training which is the training modality with the strongest evidence. This phenomenon is particularly interesting, when one considers the large number of wet-laboratories (i.e. training on animal models) being offered.

When implementing simulation-based training models, evidence-based instructional methods should be applied – including proficiency-based training, which has repeatedly been shown to positively impact learning outcomes (Gallagher et al. 2005). Thus, relevant benchmark criteria (pass/fail levels) should be defined for all training models – in select cases, automated assessment tools are available, in other cases human raters or motion-tracking tools may be used.

Relevant benchmark levels have been developed for cataract surgical training on the EyeSi simulator (Sipiteri et al. 2014; Thomsen et al. 2015). Validity of performance measurements must be verified and relevant benchmark levels should be defined for those simulation models where a benchmark level does not exist or cannot be applied due to differences in target groups. As mentioned earlier, the continual development of technology and associated changes in performance measurements, make gathering validity evidence for assessment tools an ongoing process.

Based on the findings of this thesis, we suggest that proficiency-based training on the EyeSi simulator should be implemented not only for novices but also for surgeons on an intermediate level of experience in cataract surgery (Thomsen et al. 2017).

Previous evidence suggests that variable task practice enhances learning outcomes (Spruit & Band 2014). This may correspond to training on white cataracts in addition to standard cataracts, but future studies are needed to define an appropriate level of variance for a clearly defined target group. At least, training cataract surgical skills does not seem to have an additive effect on vitreoretinal surgical performance for novices as suggested by the findings of STUDY V. This study indicates that the acquisition of intraocular surgical skills in ophthalmology appears to be domain- and task-specific, as has been previously shown for other complex procedural skills.

Implications for future research

Simulation research in ophthalmology may benefit from implementing accepted instructional methods when designing future research studies. It is also highly recommended to enrol a study population that will make research results more generalizable. From our systematic review, we noted the lack of uniformity in measurements of operative outcomes. These vary from procedural time to the ability to complete the procedure and the frequency of errors. An important area of surgical skills research is to develop a common language for the assessment of technical skills (Cox et al. 2015). We
present a flow chart that may help structure the design and interpretation of findings in future evidence-based studies in simulation-based research (Thomsen et al. 2015b).

In our systematic review, we found that most efficacy studies (65%) investigated whether repeated practice on one model leads to improved performance on the same model. In most cases, this improved performance is to be expected and these findings do not contribute significantly to this area of research. Future studies should focus on the following areas: achieving a deeper understanding of the causal relationships influencing the acquisition of technical skills in microsurgery, identification of individual factors impacting the acquisition of complex technical skills, and a comparison of various standard-setting methods to determine which method is the most appropriate for different target groups. Also, additional prospective efficacy trials are needed. Specifically, it will be important to determine the effect of simulation-based training on patient-related outcomes and provider behaviour. Lastly, the possibility of negative transfer of skills between various procedures and directionality should be studied further in future trials.

**Conclusion and Perspectives**

Training of cataract surgical skills in a virtual-reality environment is transferable to a clinical setting – also for surgeons on an intermediate experience level – but the acquired skills have not been shown to be transferable to vitreoretinal surgery performed in a simulated setting. Automated assessment provided by the EyeSi virtual-reality simulator provides useful knowledge, both for the assessment of clinical performance and as the basis for implementing proficiency-based training (given that evidence of validity has been established).

The opportunity to improve technical skills in a simulated setting prior to work in the operating room may translate into improved patient outcomes and, in turn, may have important implications for patient safety, but future studies will have to provide evidence for this link.

Future directions may include additional training possibilities for surgeons with intermediate or high levels of experience, incorporating management of complications arising during surgery and training of the entire cataract surgical procedure as an uninterrupted task in a virtual-reality environment. This would enable the interactive elements between each procedural step. The execution of every step of the procedure is becoming even more crucial because it affects the execution of the remaining steps.

The integration of simulation as a training and assessment modality in surgical specialties is not a new idea (Rehrig et al. 2008; Gillian & Saleh 2013). However, the growing evidence supporting simulation-based training and assessment should provide additional impetus for its implementation as a precursor and adjunct to clinical experience.

**References**

Accreditation Council for Graduate Medical Education (2014): ACGME Program Requirements for Graduate Medical Education in Ophthalmology. US.


Kurahashi AM, Harvey A, MacRae H, Moulton CA & Dubrowaki A (2011): Technical skill training improves the ability to learn. Surgery 149: 1–6.


a review and application to laparoscopic surgery. Psychol Res 78: 878–891.


Academic advisors: Morten la Cour, Lars Konge

This thesis has been submitted to the Graduate School of The Faculty of Health and Medical Sciences, University of Copenhagen. Submitted: 09/02/2017