Methodological shortcomings in the literature evaluating the role and applications of 3D training for surgical trainees

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Medical Teacher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>Draft</td>
</tr>
<tr>
<td>Manuscript Categories:</td>
<td>Articles</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Kostusiak, Milo; Addenbrooke's Hospital, Division of Neurosurgery; Hart, Michael; Addenbrooke's Hospital, Division of Neurosurgery Barone, Damiano; Addenbrooke's Hospital, Division of Neurosurgery Kirollos, Ramez; Addenbrooke's Hospital, Division of Neurosurgery Santarius, Thomas; Addenbrooke's Hospital, Division of Neurosurgery Trivedi, Rikin; Addenbrooke's Hospital, Division of Neurosurgery</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Surgery &lt; Discipline, Simulation &lt; Teaching &amp; Learning, Postgraduate &lt; Phase of education, Medical education research &lt; Management, Practical procedures &lt; Learning outcomes</td>
</tr>
</tbody>
</table>
Methodological shortcomings in the literature evaluating the role and applications of 3D training for surgical trainees

Milosz Kostusiak, Michael Hart, Damiano Giuseppe Barone, Ramez Kirollos, Tom Santarius & Rikin Trivedi

Division of Neurosurgery, Addenbrooke's Hospital, University of Cambridge, Cambridge UK

Corresponding author: Milosz Kostusiak, milokostusiak@doctors.org.uk, Neuroscience Department, Addenbrooke’s Hospital, Cambridge University Hospitals NHS Foundation Trust, Hills Road Cambridge CB2 0QQ
Methodological shortcomings in the literature evaluating the role and applications of 3D training for surgical trainees

Milosz Kostusiak, Michael Hart, Damiano Giuseppe Barone, Ramez Kirollos, Tom Santarius
& Rikin Trivedi

Division of Neurosurgery, Addenbrooke's Hospital, University of Cambridge, Cambridge UK

Corresponding author: Milosz Kostusiak, milokostusiak@doctors.org.uk, Neuroscience Department, Addenbrooke’s Hospital, Cambridge University Hospitals NHS Foundation Trust, Hills Road Cambridge CB2 0QQ

Running title: Shortcomings of 3D surgical training

Keywords: 3D, three-dimensional, surgery, surgical training, shortcomings, methodology
Abstract

Aim: There has been an increased interest in the use of three-dimensional (3D) technology in surgical training. We wish to appraise the methodological rigor applied to evaluating the role and applications of 3D technology in surgical training, in particular, on the validity of these models and assessment methods in simulated surgical training.

Methods: Literature search was performed using MEDLINE with the following terms: “3D”; “surgery”; and “training”. Only studies evaluating the role of 3D technology in surgical training were eligible for inclusion and assessed for the level of evidence, validity of the simulation model, and assessment method used.

Results: A total of 93 studies were analysed, and majority of reviewed articles focused on 3D displays (36) and 3D printing (35). Most of these studies were case series, the most common assessment was subjective (69), with objective assessment used by 57 studies. Very few studies provided evidence for validity of the model or the assessment methods used.

Conclusion: 3D technology has a great potential in simulated surgical training. However, the validity of this technology and strong evidence for its beneficial effects in surgical training is lacking. Further work on validation of 3D technology and assessment tools is needed.
Introduction

Apprenticeship has been a traditional surgical training model since establishment of surgery as a medical profession, where the learner follows a more experienced and skilled mentor to attain knowledge, judgement, and a range of surgical techniques (Satava et al. 2003). This model still holds true in current surgical training, however, the introduction of the European working time directive (EWTD) and relative increase of surgical trainees to core surgical procedures have led to significant reduction in the hours spend in the operating theatre (Breen et al. 2013; Toll and Davis 2010). Consequently, these changes have necessitated the search for alternative training methods, mainly in the form of simulation.

Simulation is a crucial part of surgical training as it emulates a real life scenario that allows a trainee to practice procedural skills in a safe environment (Sutherland et al. 2006). It has been widely adopted throughout the health care profession and can range from low-fidelity (e.g. glove model for suturing) to high-fidelity simulation models (e.g. pig model for laparoscopic cholecystectomy) (Van Bruwaene et al. 2015). Currently, there is a plethora of simulation models available for the trainees, and they can be categorised into physical, cadaveric, animal, harvested tissue, computer simulation and hybrid models (Samia et al. 2013). Human cadavers are incorporated into surgical training throughout the UK and other countries, however, the evidence for effectiveness of cadaveric workshops in higher surgical training is limited (Gilbody et al. 2011). Furthermore, due to a limited supply of cadavers and the cost associated with it, there has been a trend towards using more affordable and sustainable models, for example three-dimensional synthetic models (Hochman et al. 2014). Three dimensional (3D) technology has improved immensely over the recent years and led to the development of 3D virtual reality (VR), printing, and interactive mobile applications, all of which are slowly being integrated into current surgical training (Lewis et al. 2014; Olivieri et
al. 2016). However, the question remains whether 3D technology in simulation training has any beneficial effect on surgical skill acquisition.

Surgical performance, competence, and skill acquisition can be measured through various assessment methods, broadly categorised into formative and summative, where former is aimed at providing feedback on the performance and latter is used to formally assess the acquisition of the skills required for completion of the training (Evgeniou 2013). The majority of assessment tools in surgical training are subjective and most of these tools are checklists (Alkhayal et al. 2012), for example the Structured Technical Skills Assessment Form (STSAF) (Winckel et al. 1994). These tools require standardised training of the assessors in order to provide reliable results and reduce possibility of a rater bias. On the contrary, objective assessment tools do not involve subjectivity of the assessor and thus are more reliable. Examples of such tools include end product assessment, virtual reality (VR), hand motion analysis (HMA) and objective structured assessment of technical skills (OSATS) (Moorthy et al. 2003). VR and HMA generate objective results, such as task completion times, or numbers of inaccurate or corrective manoeuvres, which can be used to assess the performance by comparing the results against ‘expert controls’ (Chin et al. 2011; Perrenot et al. 2012). However, OSATS, which uses checklists with performance being assessed by expert surgeons, has shown high reliability and construct validity, which suggests that with extensive validation processes, these checklists can be used as objective assessment tools (Reznick et al. 1997). OSATS still remains the most popular assessment tool for open and laparoscopic skills, but unfortunately the transition from bench or cadaveric simulation models to assessment in operating theatres is seldom seen (Shaharan and Neary 2014).

This review focuses on the methodological shortcomings in evaluating the role and applications of 3D technology in surgical training in particular, on the validity of these models and assessment methods in simulated surgical training.
Methods

An extensive literature search was performed using the MEDLINE search engine. The search was performed using the following terms: “3D”, “surgery” and “training”. Only the studies evaluating the role of 3D technology in surgical training, and including both simulation and real-time operations, were accepted. Selected studies were then assessed for the level of evidence, setting and training modality, whether the model was validated for the purpose of the surgical training, assessment methods used, and the overall outcome.

Results

The literature search identified 761 unique studies that had evaluated the use of 3D in simulated surgical training. After careful screening of titles and abstracts, 93 studies met the inclusion criteria and the full texts of these were then analysed. The studies were screened for the type of research, setting and training modality, assessment method and overall outcome. The vast majority of the 3D related studies included 3D displays (36, 38.7%) and 3D printing (35, 37.6%), the remaining studies used VR, 3D images and videos as well as 3D navigation (Table 1). 3D displays were used with laparoscopic and endoscopic equipment. 89 studies (95.6%) were carried out in a non-clinical setting, whereas the rest were performed in the operating theatre.

Levels of evidence of the studies were assessed according to the widely cited and published criteria developed by the Oxford Centre for Evidence-based Medicine. The level of evidence of these studies ranged from 2b (e.g. poorly designed randomised clinical trials) to 5 (e.g. an expert opinion). The majority of the studies were case series (level 4 evidence) (Table 1).

Validity is a concept defined as “the property of being true, correct, and in conformity with reality”, and includes face, content, construct, concurrent, discriminative and predictive
validity (Gallagher et al. 2003). These can evaluate different aspects of the intervention, for example, while concurrent validity relates to correlation to the gold standard simulator, the construct validity indicates that the simulator is able to distinguish between different levels of competence (McDougall 2007). Regrettably, very few studies attempted providing evidence for validity of the model. 15 studies had reported face validity, 8 had reported content validity, 3 reported construct validity, and 67 did not report any validation for their models. However, it is important to note that some of the studies used established 3D models or cadaveric set-ups that had been previously validated.

Assessment methods were divided into objective and subjective assessment categories. Several studies utilised both methods, however only 57 studies (62%) have used objective assessment tools. These tools varied considerably, ranging from an anatomy quiz to measuring performance and quality of the end product. Subjective assessment tools were used by only 69 studies (74.1%) and included questionnaires or rating scales. The results showed that 3D techniques were found to be superior to 2D in 90 studies (96.7%), and participants from all of the studies perceived 3D as useful or very useful. Unfortunately, only a small fraction of the 3D studies had used validated assessment methods. In one study, laparoscopic performance was measured using a validated Fundamentals of Laparoscopic Surgery (FLS) skills set, which measures time to completion, number of attempts required to achieve proficiency, and the number of errors. Participants in 3D had performed better and completed the task quicker with fewer errors, and required fewer repetitions to achieve proficiency (Tanagho et al. 2012). Similar results were obtained by using Imperial College Surgical Assessment Device (ICSAD) to measure the precision of 3D endoscopic performance (Taffinder et al. 1999). In another study, 3D printed models improved anatomical knowledge as evidenced by a post-intervention anatomy quiz, which is also an objective measure (Ng et al. 2015). [Table 1 near here]
Discussion

This review provided an overview of the differing methodologies and methodological strategies being used to evaluate the effectiveness or value of 3D technology in simulated surgical training. Before the introduction of the EWTD, a trainee would be exposed to an average of 32,000 hours of experience, compared to almost half of that time after the introduction of the EWTD (Cresswell et al. 2010), thus making simulation an invaluable tool for surgeons in training. However, the effectiveness of simulated training depends on several aspects, including curriculum, validity of the simulation model and assessment methods. This review has presented potential shortcomings when evaluating the role and application of 3D technology in surgical training.

Quality of the evidence

As with any medical research, well designed and performed studies would produce better quality results and thus have higher impact. This is also true for surgical education research, but arguably it is much harder to perform well-designed studies. This review has shown that the evidence in the literature assessing 3D intervention in surgical training is of low to moderate quality. The highest level of evidence was reached by only few studies that have used blinded randomisation, and several other studies have performed randomised trials, but without using a blinded assessor. Majority of the studies were case series and in some studies the methodology was not clear enough to determine the class of evidence, and in these cases they were automatically assigned as level 4 evidence. There are several possible limiting factors that affect the design of the study and these include difficulties in randomisation, time constraints, limited availability of the simulation models and lack of validation of simulation
models or assessment methods, and all of these should be addressed to produce high quality evidence.

**Validation of simulated models**

In the reviewed literature, only a handful of studies had reported face or content validity for their models. Due to the fact that most of the participants in those studies were either students or junior trainees with little previous exposure to clinical environment, it is unwise to assess their performances and the usability of the models without prior validation. Inexperienced participants are more likely to give higher scores to 3D printing models as they would not be able to compare the model to the real life anatomy. Therefore, simulation models must resemble the real scenario as close as possible for the developed skills to be transferable to the clinical world, which can be achieved through the process of validation, and thus can be used effectively for the skill assessment purposes.

**Assessment methods in surgical simulation**

Since simulation has become an integral part of the surgical training, a plethora of assessment methods were created to provide the trainee with formative and summative feedback. These assessment tools must be valid, reliable, feasible, in order to have an educational impact. In addition to ‘model validation’, assessment tools should ideally contain the essential elements of validity, including construct, content, and predictive validity. It is also crucial that these assessment tools provide the participants with accurate feedback so they can monitor skill acquisition and improve their surgical abilities. To date, this has been best achieved through objective assessment methods, however, it is also evident that subjective assessment tools have been successfully employed for that reason. Studies referenced in this review have utilised an extensive array of assessment tools, ranging from subjective questionnaires to
objective quantitative data collection. In nearly all studies, the participants have rated the 3D training models as ‘useful’ for surgical training in post-intervention questionnaire. User’s satisfaction is an important aspect of model evaluation, but are subjective in nature, and expert panels or consensus groups that determine the characteristics of good or ‘model’ simulations are potentially liable to similar criticisms. Therefore, many have focussed on obtaining quantitative data, such as the time to complete task, error rate or number of corrective manoeuvres required. These parameters are very generic and can be used to assess any surgical procedure, and thus were the most commonly used assessment tools to obtain quantitative data. Fundamentals of Laparoscopic Surgery (FLS), McGill Inanimate System for Training and Evaluation of Laparoscopic Skill (MISTELS) (Fried et al. 2004), or task specific checklists and global rating scales (GRS) (Ezra et al. 2009) are some of the examples of validated assessment tools used only by handful of reviewed studies.

**Study design and its limitations**

The study design is single most important step when setting up a study to evaluate a new intervention. However, the highest class of evidence is not always achievable or feasible, thus one must carefully chose the most suitable methodology. When assessing the effectiveness of 3D simulation models in surgical training, the authors would need to consider several aspects, including, but not limited to, characteristics of the participants, study environment, validation of the simulation model and assessment methods.

The majority of studies in surgical education employs medical students, core and higher surgical trainees, and consultants as participants. This inhomogeneity among participants allows for the intervention to be tested at different stages of surgical training and might be useful at assessing construct validity. However, varying experience levels within the cohort
may prevent the study to be conducted as a well-designed randomised controlled trial (RCT) due to possible selection bias leading to baseline imbalances (Roberts and Torgerson 1999). Furthermore, trainees may refuse randomisation as they would not want to miss critical educational experiences, or due to clinical rotation differences the trainees may experience intervention at different time points (Sullivan 2011b). In addition, the control group within RCT that evaluate 3D technology would be exposed to the current gold standard, 2D equivalent, and thus blinding of participants in this example would not be possible. This may lead to comparison choice bias whereby the control group pre-emptively believes 3D to be better and rating 2D lower (Jadad and Enkin). One of the ways to reduce the risk of selection and control biases would be to apply a single subject design, where participants are equally exposed to all intervention groups, therefore the performance in each intervention can be compared within subject. This study design removes the need to randomise participants, each participant serves as its own control, and fewer participants may be required. However, as with all crossover-design studies, there is a risk of carry-over or order effects, where the effect of the first intervention is carried onto the next. The way to minimise this risk, is to incorporate a lengthy washout period, i.e. the time between the interventions. However, this would not necessarily negate the impact of the sequence of interventions. This type of design is the most commonly reported in submissions to medical education journals (Sullivan 2011a).

Another important aspect of study design is the selection of the best assessment (or evaluation) method(s). The choice of the assessment tool must be governed by the research question, study design, type of intervention and participant characteristics. Nonetheless, one must critically consider the feasibility of the assessment tool, for example, assessing temporal bone surgery skills using virtual reality or 3D printed models is more practical and safe compared to real-life surgery (Linke et al. 2013; Mowry et al. 2015). The matter becomes
more complicated, however, when evaluating the intervention that is being tested on a large
group. In the case of assessing 3D imaging in surgical anatomy lectures, where the whole
class may participate in the study, self-assessment may offer a feasible alternative. Self-
assessment of surgical skills has been shown to have good reliability and validity
(MacDonald et al. 2003; Mandel et al. 2005), but its role in assessing anatomy knowledge is
less clear (Anthoney 1986). Confidence-based marking (CBM) and self-assessment were
shown to have a strong correlation with test performance (Barr and Burke 2013; Schiekirka et
al. 2013), while other forms of self-assessment had weaker correlations (Vivekananda-
Schmidt et al. 2007). These pre- and post-intervention questionnaires, such as the Likert
scales, are very common in surgical education research, but can be subjective. Thus, a more
suitable and objective tool to evaluate the effectiveness of 3D imaging in anatomy teaching
would be pre- and post-intervention tests (Muller-Stich et al. 2013).

Future Directions

There is limited evidence for the effectiveness of 3D teaching of neuroanatomy in surgical
education. To this end, there is a need to construct a study that aims to assess whether 3D
visualisation of neuroanatomy (3D lectures or cadaveric dissection) is an effective method of
enhancing the knowledge and confidence of neurosurgeons, and how it compares to
traditional training.

Accepting that any study proposal will have some limitations, we propose a study that would
employ a three-armed experimental design where participants are exposed to either standard
training (2D lecture based visualisation and cadaveric training) or one of two novel
interventions: 3D lecture based visualisation alone or 3D lecture based visualisation
combined with cadaveric training. This design would allow a comparison of several different
interventions at the same time. The study group should include participants of varying levels
of experience, including undergraduate medical students, surgical trainees and consultants. However, the majority of the participants are likely to be neurosurgical trainees as this intervention is primarily aimed at them. Furthermore, both medical students, especially at the very early stages of education, and consultants will serve as negative and positive controls, which would allow a measure of construct validity of the tested model.

The assessment tool(s) will include both objective and subjective methods. Baseline assessment (pre-intervention) and final assessment (post-intervention) will be employed and these may include pre- and post-intervention quiz, combined with self-assessment. This would allow for the participants to be act as their own controls. For cadaveric training and assessment, the use of a validated task specific checklist would be used, and an experienced surgeon should carry this assessment out. Furthermore, a comparative survey may be used if the participants have been exposed to both interventions (e.g. 2D vs 3D lectures). The overall effectiveness or value of these models will be assessed using comparative statistical evaluations of the self-reporting rating scales and observer checklists, between groups and within groups with pre- and post- comparisons. The reliability of self-reporting compared with independent observer checklists will be determined by correlation analyses.

Conclusion

3D technology has a great potential in simulated surgical training as the majority of trainees find the concept very useful and believe their surgical skills improve after exposure to such models. However, there are a variety of methodologies used to determine the effectiveness of 3D simulation, and the question still remains whether these models have a beneficial effect on either skills acquisition or surgical performance. Lack of model validation, poor quality of study design and inadequate assessment methods are the main methodological shortcomings and whilst not all of these shortcomings are likely to be addressed in any single study, the
proposed study design offers a pragmatic way of implementing a comparative assessment paradigm into existing 3D assisted training courses.

**Practice points:**

1. Simulation training forms an integral part of surgical trainee’s pathway.
2. 3D technology can be used effectively in simulated surgical training.
3. Studies assessing 3D technology in surgical training must carefully consider appropriate study design with validated assessment methods.

**Notes on contributors**

All authors are part of Neuroanatomy Education Group at Division of Neurosurgery, Neurosciences Department, Addenbrooke’s Hospital, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK.

**Declaration of interest:** None
References


Evgeniou E. 2013. Assessment methods in surgical training in the united kingdom.


Jadad AR, Enkin MW. Bias in randomized controlled trials.


Table 1. Twelve most important studies identified in the literature review

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Modality</th>
<th>Setting</th>
<th>n</th>
<th>Objective Evaluation</th>
<th>Subjective Evaluation</th>
<th>Evidence Level</th>
<th>Methodology validation</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Alaraimi et al. 2014)</td>
<td>3D Display</td>
<td>Lab</td>
<td>50</td>
<td>Time and error rate</td>
<td>N/A</td>
<td>2b</td>
<td>Face validation</td>
<td>Improved accuracy, but no effect on global performance</td>
</tr>
<tr>
<td>(Tanagho et al. 2012)</td>
<td>3D Display</td>
<td>Lab</td>
<td>33</td>
<td>Time, number of repetitions and error rate</td>
<td>Survey</td>
<td>4</td>
<td>Validated assessment method (FLS)</td>
<td>Improved performance and preferred by participants</td>
</tr>
<tr>
<td>(Curro et al. 2015)</td>
<td>3D Display</td>
<td>Theatre</td>
<td>80</td>
<td>Time</td>
<td>Survey</td>
<td>2b</td>
<td>Yes (real patients)</td>
<td>3D not superior to 2D equivalent</td>
</tr>
<tr>
<td>(Felisati et al. 2013)</td>
<td>3D Display</td>
<td>Theatre</td>
<td>8</td>
<td>Post-operative follow up</td>
<td>Survey</td>
<td>4</td>
<td>Yes (real patients)</td>
<td>Improved performance and depth perception</td>
</tr>
<tr>
<td>(Engel et al. 2015)</td>
<td>3D Printing</td>
<td>Lab</td>
<td>19</td>
<td>Time and end product</td>
<td>N/A</td>
<td>4</td>
<td>Validated model and checklist</td>
<td>Validation of the model</td>
</tr>
<tr>
<td>(Lin et al. 2014)</td>
<td>3D Printing</td>
<td>Lab</td>
<td>25</td>
<td>Time, the maximal acceleration of drill</td>
<td>Survey</td>
<td>3b</td>
<td>Construct, face and transfer validity of the model</td>
<td>Validation of the model</td>
</tr>
<tr>
<td>(Ng et al. 2015)</td>
<td>3D Printing</td>
<td>Lab</td>
<td>72</td>
<td>Quiz</td>
<td>Survey</td>
<td>2b</td>
<td>No</td>
<td>Higher quiz scores and preferred by participants</td>
</tr>
<tr>
<td>(Chalas)</td>
<td>3D VR</td>
<td>Lab</td>
<td>26</td>
<td>Time and error rate</td>
<td>Survey</td>
<td>4</td>
<td>Face, content</td>
<td>Validation</td>
</tr>
<tr>
<td>Study</td>
<td>Setting</td>
<td>Accuracy</td>
<td>Format</td>
<td>Control</td>
<td>Score</td>
<td>Preferred by Participants</td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------------</td>
<td>---------------</td>
<td>----------</td>
<td>---------</td>
<td>---------------</td>
<td>-------</td>
<td>---------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ani et al. (2011)</td>
<td>3D VR Lab</td>
<td>89</td>
<td>Quiz</td>
<td>N/A</td>
<td>3b</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Donnelly et al. 2009)</td>
<td>3D Images Lab</td>
<td>73</td>
<td>Quiz</td>
<td>N/A</td>
<td>2b</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Metzler et al. 2012)</td>
<td>3D Images Lab</td>
<td>73</td>
<td>Quiz</td>
<td>Survey</td>
<td>2b</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Drapkin et al. 2015)</td>
<td>3D Video Lab</td>
<td>17</td>
<td>Quiz</td>
<td>Survey</td>
<td>2b</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Prinz et al. 2005)</td>
<td>3D Images Lab</td>
<td>73</td>
<td>Quiz</td>
<td>Survey</td>
<td>2b</td>
<td>No</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>