Spring 2016

Objective measures of operating room wire navigation performance

Leah Kristine Taylor
University of Iowa

OBJECTIVE MEASURES OF OPERATING ROOM WIRE NAVIGATION PERFORMANCE

by

Leah Kristine Taylor

A thesis submitted in partial fulfillment
of the requirements for the Master of Science
degree in Biomedical Engineering in the
Graduate College of
The University of Iowa

May 2016

Thesis Supervisors: Associate Professor Donald D. Anderson
Professor Geb W. Thomas

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

_____

MASTER'S THESIS

_____

This is to certify that the Master's thesis of

Leah Kristine Taylor

has been approved by the Examining Committee for
the thesis requirement for the Master of Science degree
in Biomedical Engineering at the May 2016 graduation.

Thesis Committee: _____
Donald D. Anderson, Thesis Supervisor

_____
Geb W. Thomas, Thesis Supervisor

_____
David G. Wilder

_____
Matthew D. Karam

Acknowledgements

I would like to thank my thesis supervisors Drs. Don Anderson and Geb Thomas for their guidance and support over the past two years. Having professional role models who demonstrate success through example made this achievement possible. Collaboration from orthopedic residents, staff, and surgeons, especially Dr. Matthew Karam with his clinical vision and leadership, was essential to the success of this research. Additionally, I would like to recognize Dr. David Wilder for providing academic support and insight on my committee. I would also like to thank faculty and fellow students in the Orthopedic Biomechanics Lab, whose daily support was indispensable. Finally, I owe the most gratitude to my family; my parents who champion my personal and professional development unconditionally and my husband for always pushing me.

This work was partially supported by grants from the American Board of Orthopedic Surgery as well as the Agency for Healthcare Research and Quality.

Abstract

Surgical education is changing with a push for the inclusion of simulated environments for training and assessment. A critical element of this transition is linking performance in a lab based setting with the actual operating room. Currently there is no widely accepted, reliable tool for measuring surgical skill in the operating room. Ubiquitous video and imaging technology provide unique opportunities to develop metrics to meet this need. Hip fracture surgery is a promising area in which to develop these measures because hip fractures are common, the surgery is used as a milestone for residents, and it demands technical proficiency.

Resident surgeons wore a head-mounted video camera while performing surgical open reduction and internal fixation of hip fractures using a dynamic hip screw or telescoping screw plate. The wire navigation portion of the video was analyzed. Data collected from the video included: duration of wire navigation, number of fluoroscopic images acquired, and the degree of intervention by the surgeon's supervisor. To determine the reliability of these measurements, four independent raters performed them for two cases. Ten raters independently measured the tip-apex distance (TAD), which reflects the accuracy of the surgical placement of the wire, on 7 cases. These metrics for 15 cases were then compared to experience metrics including point in residency and number of previous cases performed. A composite performance score was computed by summing the average standardized values of the four performance metrics. Expert surgeon opinion, the Objective Structured Assessment of Technical Skills (OSATS) score of two traumatologists, was compared with these metrics.

The inter-rater reliability analysis for all video-based measures produced a Cronbach's Alpha of 0.99 and for the combined TAD measurements a Cronbach's Alpha of 0.97. There was significant correlation between surgical experience and both procedure duration and tip-apex distance. The composite performance metric significantly correlated to both weeks into residency -0.55 ($p=0.03$) and cases logged -0.66 ($p=0.01$).The OSATS score was only significantly correlated to surgery duration and number of fluoroscopic images.

Several of the video-based metrics and TAD measurement were consistent across the raters and are useful for performance assessment. The wire navigation performance metrics, time and TAD, were shown to differentiate surgical experience. A composite score incorporating multiple performance metrics also provided strong correlations with surgical experience. The methods presented have the potential for truly objective assessment of resident technical performance in the operating room, a critical step towards competency based education.

Public Abstract

There is no widely accepted tool to assess an orthopedic surgeon's technical skill in the operating room. With changes in surgical education, simulators are being investigated for learning and assessing technical skills, but a link between the actual operating room is needed to ensure they are effective. Hip fracture surgery is a good starting point to develop these measures because hip fractures are common and fixation is a difficult task.

Resident orthopedic surgeons wore a head-mounted video camera during hip fracture surgery. Data collected included: duration, number of x-ray images, the supervising surgeon intervention, and tip-apex distance (TAD, a measure of how accurate the implant is placed). To determine the reliability of these measurements, four raters performed them for two cases. Ten raters measured the tip-apex distance (TAD) on 7 cases. These performance metrics for 15 cases were compared to experience of the residents, both point in residency and number of previous cases. A composite performance score was computed using the four metrics. The metrics were also compared to two practicing surgeons' assessment of skill.

The inter-rater reliability of the performance metrics was high (0.97-0.99) showing these measures are consistent between different raters and useful for assessment. There was a significant relationship between resident experience and the metrics of duration and TAD. Expert opinion was related to duration.

These metrics provide objective assessment of resident technical performance in the operating room by a non-expert, an important step towards competency based education. Their validity is shown with correlation to surgical experience.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The world of surgical education is moving away from the traditional apprenticeship model. In 2003, the Accreditation Council for Graduate Medical Education (ACGME) established a restriction of 80 duty hours per week for residents. An adverse effect of this enactment was a decrease in resident's time spent absorbing skill and technique in the hands-on environment of the operating room. Partially due to this change, the current method of surgical training was brought into question by the government, insurance companies, and the general public.[1] In 2009 and 2013 articles were published in media giants such as The New York Times suggesting a degree of incompetence of graduating residents.[2] With a decrease in exposure to procedures, a new system of verifying proficiency and establishing confidence in graduates' competence is vital. Another consequence of the current time-based residency program is an unavoidable gap between surgical skills of various residents depending on the cases they are presented. The perceived skill of an individual resident can determine exposure to cases; a more talented resident may be given more opportunities to perform challenging procedures further increasing the inter-resident skill gap. The need for a consistent method of training and credentialing board certified surgeons to an established level of competency is at the forefront of medical governing bodies focus.

Some specialties, such as general surgery, have made advances in this evolution; however orthopedic surgery is lagging behind. A competency based orthopedic surgery program would instill confidence in graduating residents' ability and a level of added

patient safety with residents only being permitted to operate after achieving a determined competence. A proof-of-principle study began in 2009 at the University of Toronto; a completely competency-based residency was employed alongside a traditional time-based program. Those residents enrolled in the competency based program outperformed peers in addition to more senior residents.[3] The critical first step in transitioning to a competency centered education system is developing assessments that accurately and reliably demonstrate technical skill. Graduate medical education committees have failed to establish standard measures of competency.[4] Simulators have been introduced to residency programs as educational and assessment tools. The merit of simulators in providing a low risk environment for residents to practice technical skills, while not jeopardizing patient safety, is evident. For simulators to be utilized, skill transfer from the simulated setting to the OR is vital [5] and this essential connection is identified as the focus of future work.[6; 7] The first step in establishing skill transfer for prediction and improvement purposes is being able to reliably assess technical performance in the high stakes operating room environment. Once performance in the OR can be assessed, standards of acceptable performance can be set and the credentialing process reformed.

The primary objective of this research is to address the difficult task of objectively assessing operating room surgical skill in orthopedic trauma surgery. In the United States between 1986 and 2005 hip fracture incidence was 957.3 per 100,000 women and 414.4 per 100,000 men.[8] Not only are these fractures common, but healthcare costs and mortality for hip fractures are greater than all other osteoporotic fractures combined. In 2005 hip fractures were estimated to account for 72% of health care costs related to

fractures. [9] Fixation of intertrochanteric hip fractures has been identified as a critical task to be mastered prior to graduation from residency.[6] The task chosen, wire navigation, is used in a variety of orthopedic procedures including hip fracture fixation. Wire navigation plays a crucial role in the success of fixation of intertrochanteric hip fractures with a plate and screw fixation device such as a Dynamic Hip Screw (DHS) or Telescoping Screw Plate (TSP). The wire serves as a guide and determines the placement of the final implant, which is linked to the clinical success of the fixation.

The second chapter addresses the development of objective measures of operating room performance that can be reliably made by non-experts. Video recordings from a head mounted camera (Figure 1.i) and intraoperative images are captured and used for metric evaluation. Analysis of inter- and intra-rater reliability demonstrates these metrics are reliable and provide a promising platform for future assessment applications. A detailed account of the elements of this innovative methodology is described. This manuscript was accepted for publication in the Journal of Surgical Education[10] and per the author permissions can be included in full in a thesis (https://www.elsevier.com/about/companyinformation/policies/copyright/permissions).

Figure 1.i Intraoperative image of a GoPro camera prior to incision.

The third chapter uses the metrics presented in Chapter 2 to assess orthopedic resident performance on 15 cases in the operating room and correlates these to surgical experience. Surgeries at the University of Iowa were included with collaboration of orthopedic staff and resident surgeons. Experience was measured from previous case load and time spent in residency. Additionally, two attending traumatologists rated the recordings to provide a parallel comparison of expert assessment of the same performance. Significant correlation between surgeon experience and the metrics of time and tip-apex distance (TAD) was found. Although the case size is small, these findings provide the groundwork for additional institutions to collect data across residency programs and provide necessary information to credentialing organizations. This chapter will be submitted to Clinical Orthopaedics and Related Research.

Chapter 2: Establishing Metrics – Assessing Wire Navigation Performance in the Operating Room

Leah Taylor, BS,[a,b] Geb W. Thomas, PhD,[a,c] Matthew D. Karam, MD,[a]
Clarence Kreiter, PhD,[d] Donald D. Anderson, PhD,[a,b,c]

[a]Department of Orthopedics & Rehabilitation
[b]Department of Biomedical Engineering
[c]Department of Mechanical & Industrial Engineering
[d]Department of Family Medicine

The University of Iowa
Iowa City, IA, USA

2.1 Introduction

An orthopedic surgeon must pass a written examination, an oral examination, and complete a 5-year residency program at an accredited institution to become board-certified in the United States. Case logs, required by the Accreditation Council for Graduate Medical Education, track the number and level of participation in specific procedures and are used to establish competence, but they fall short of measuring technical skill.[11] Only 31% of orthopedic resident leaders feel that the case log volume should play a role in credentialing.[12] It is unclear what might replace case logs, but the president of the American Orthopaedic Association have recently argued for competency-based training.[2] A program focusing on competency-based end points offers advantages over a time-based system, including increased efficiency for learners and opportunities for focused instruction on problem areas.[13] A crucial step toward a competency-based

program is identification of objective, quantifiable measures for assessing surgical performance.

Presently there is no widely accepted tool in clinical practice for efficiently and reliably measuring technical surgical skill.[14] The Objective Structured Assessment of Technical Skill (OSATS) is a reliable and valid tool that has been increasingly used in orthopedic skills training. It uses a global rating approach to structure expert evaluation of technical skills, with the experts working from a list of specific operative competencies that are each rated on a 5-point Likert scale anchored by behavioral descriptors. The OSATS is widely accepted as the "gold standard" for formative surgical skill assessment, at least partly because of the significant correlation between OSATS scores and postgraduate residency year.[15] OSATS is not recommended for summative evaluation.

OSATS scoring does have several important limitations including: potential personal bias [16], the inconvenience and expense of using staff surgeon expert for grading, and the lack of objective demarcations between acceptable and unacceptable performance.[1] Furthermore, a recent study showed that OSATS scoring methods do not effectively assess the quality of the surgical result, suggesting that efforts must be made to incorporate assessment metrics that reflect the quality of the surgical result.[17] A more systematic and efficient approach might include a controlled simulation with truly objective scoring metrics. The Fundamentals of Laparoscopic Surgery trainer, a hands-on skill examination required for certification by the American Board of Surgery, is one such example of this approach. The examination measures technical competence of

laparoscopic surgical skills by grading the efficiency and precision of five tasks executed on a box trainer.[18]

Other simulators provide quantifiable performance metrics that may eventually be useful in measuring operating room (OR) performance. For example, a promising set of visual parameters: prevalence of instrument loss, triangulation time, and prevalence of lookdowns, appear to correlate with other validated skill assessments, including a global rating scale and motion analysis.[19] Motion tracking may also one day prove to be useful in the OR. The Imperial College surgical assessment device (ICSAD) uses electromagnetic tracking to collect motion data and measure metrics such as number and speed of hand movements, distance traveled, and overall time.[20]

Collecting videos of a surgery presents a more immediate approach to measuring OR performance. To relate surgical technical skill to outcome, Birkmeyer et al. used blinded video recordings of bariatric surgeries. A video recording of an operation was graded in terms of technical skill by blinded surgeons, and this score was correlated to postoperative complication rate. Lower skill was associated with an increase in operation time, as well as with rate of reoperation and complication.[21] In a study by Beard et al., two different surgeons assessed direct observation in the OR completed by a third surgeon and video recordings of the same operations.[22] Inter-rater reliability between the three assessing surgeons was 0.96, suggesting that with the elimination of bias, it is possible to use video recorded assessments for summative assessments.

Although there are not yet any studies of measuring *orthopedic* surgical skill with video recordings, hip fracture surgeries are a good place to start because they are

common and they demand technical proficiency in wire navigation, a generalizable skill. The Accreditation Council for Graduate Medical Education requires that a minimum of 30 cases be logged during an orthopedic residency, suggesting that hip fracture surgery is already considered to be a foundational part of an orthopedic surgeon's skill.[23] Wire navigation is a challenging task involving the insertion of a stainless steel pin (K-wire) along a specific trajectory to a position in a bone; it can also serve as a guide for surgical implant placement. This skill, in some form or another, is a fundamental aspect of 7 of the 16 orthopedic milestone cases, with a procedure's success depending on the path and final position of the wire. In a simulated wire navigation task, experienced surgeons more accurately place the guide wire in less time than novice residents.[24]

Wire navigation used in the treatment of inter-trochanteric fractures of the hip with a dynamic hip screw (DHS) presents a particularly good initial target for OR performance assessment, because the success of the surgery depends upon a straightforward measure – the tip-apex distance. The tip-apex distance (TAD) is the sum of the distances from the tip of the hip screw to the apex of the femoral head on the antero-posterior (AP) and lateral radiographs. The sum does not represent the physical three-dimensional distance from the wire to femoral head apex, but rather the addition of two two-dimensional measurements on the orthogonal fluoroscopic images (Figure 2.i).[25] A TAD value of 25 mm or less significantly decreases the risk of screw cut out; therefore, surgeons placing a DHS seek to achieve a center-center position with a TAD of less than 25 mm.[26] Although TAD provides an unambiguous measure of surgical placement, OR performance must also account for how a surgeon balances precision

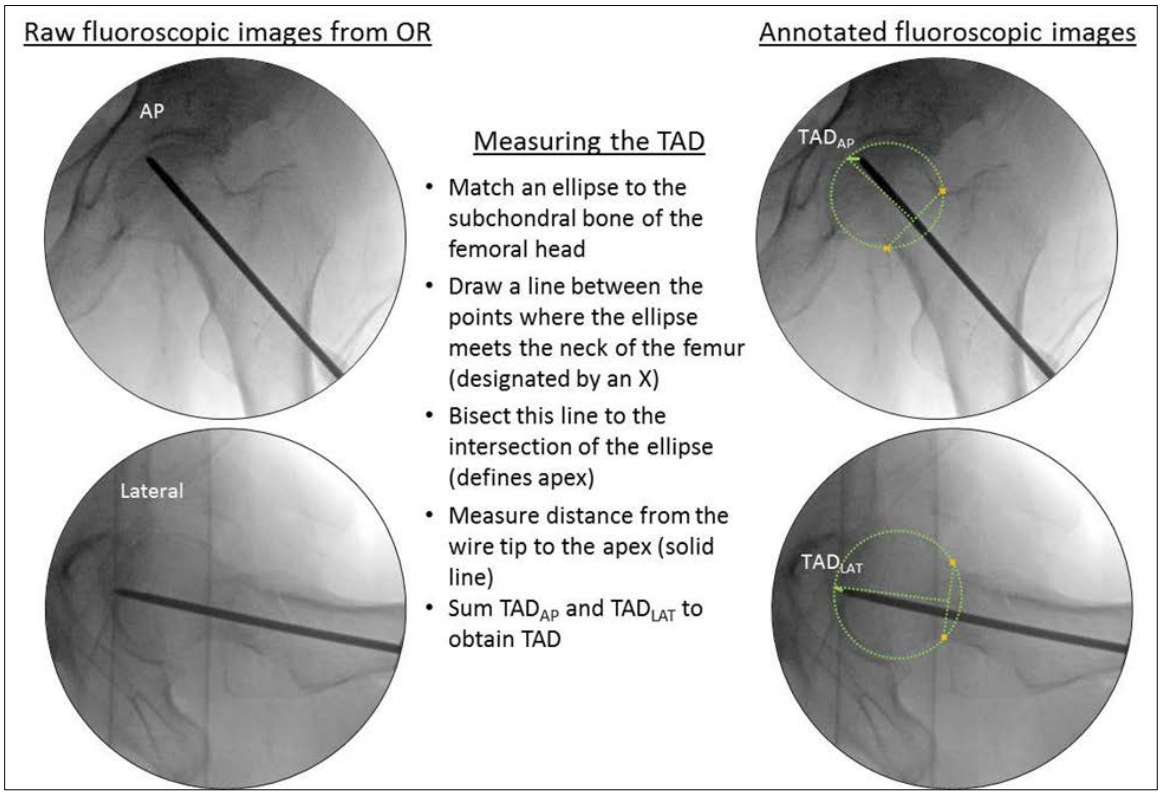
against surgery duration. In the present situation, the K-wire serves as a guide for the placement of the implant. The surgical placement goal for the K-wire is effectively equivalent to that of the implant placement. Consequently, the K-wire TAD is a convenient proxy for the surgeon's skill in achieving a desirable implant TAD.

Operation time correlates with both cost and patient outcomes. A surgeon's experience and competence tends to reduce operating time.[27] An operation time of greater than 240 minutes carries a significantly higher risk of surgical site infection in hip fractures.[28] Surgical performance in orthopedics also depends on the effective use of fluoroscopic imaging. Intraoperative fluoroscopy is crucial to visualization during many orthopedic procedures but is also a source of radiation exposure to the patient and OR staff. Adverse effects of radiation exposure due to scatter during C-arm fluoroscopic imaging include an increased risk of cancer, sterility, and eye damage.[29-31] Kahn et al. explored radiation used by residents in DHS procedures and showed the primary differentiator in radiation exposure is surgical experience.[32] Multiple studies have also shown that an increase in surgical experience results in a decrease of radiation exposure.[29; 33-35]

Another measure of surgical performance that considers how meticulous and efficient the surgeon is at exposing the patient to the least possible insult or injury is the number of holes created in the bone cortex. When a wire does not have an appropriate trajectory, it must be removed then reinserted, which increases surgical time and radiation exposure. If the wire is advanced too far it may breach the femoral head and disrupt the joint space.

The principal objective of this study was to quantitatively assess wire navigation skill in the OR. The specific hypothesis is that the tip-apex distance, the surgical duration, the number of fluoroscopic shots, and the number of bone cortex breaches may be consistently measured using augmented video recording of the procedure from a head-mounted camera.

Figure 2.i Measuring the TAD.



2.2 Methods

Surgeries involving open reduction and internal fixation of intertrochanteric hip fractures performed using a DHS were identified and enrolled in the study. Patient and resident informed consent were obtained following protocol approved by the Institutional Review Board at the University of Iowa, IRB # 201409755. The surgeries were

completed at the University of Iowa Hospitals and Clinics by the on call Orthopedic Trauma team. Patient mean age was 83.7 years (standard deviation 5.6 years), BMI 28.2 (standard deviation 4.1), and 71.4% (5 of 7) were female. One of the operating residents was a PGY-2 (second year of postgraduate training), five were PGY-3, and one was a PGY-5. There were five different supervising surgeons (two staff surgeons and three senior level residents) for the seven cases. All raters in the study were biomedical engineering graduate students working in the Orthopedic Biomechanics Laboratory at the University of Iowa.

Resident surgeons wore a GoPro Hero3 (GoPro, San Mateo, CA) head-mounted video camera during the surgery (Figure 2.ii). Recording started before the operating resident entered the sterile field. All intraoperative fluoroscopic images were automatically saved after each exposure for later evaluation. The following data were collected by a researcher who viewed the video and the saved fluoroscopic images from that portion of the surgery involving the wire navigation procedure, beginning with the first fluoroscope image with the guide wire and ending with the final fluoroscope image with the wire ready for the implant: duration of wire navigation procedure, number of fluoroscopic images collected, number of pullbacks, number of cortex head breaches, number of times the C-arm moved between AP and lateral imaging positions, and the degree of intervention of the surgeon's supervisor**.** Table 2.i provides a summary of the data source for each of these metrics.

Figure 2.ii Intraoperative picture of a resident operating (left) and camera view during wire navigation (right).

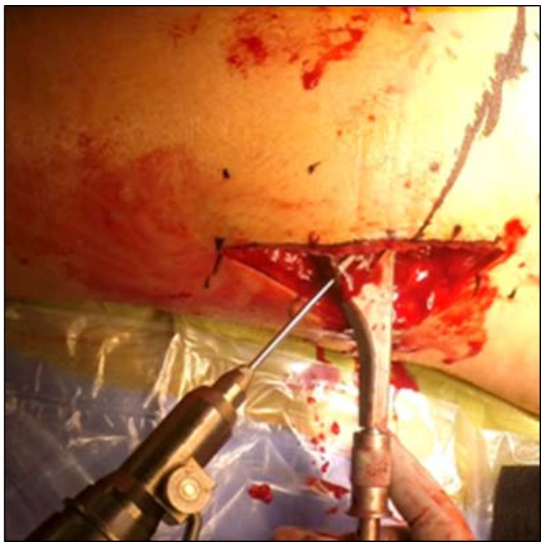


Table 2.i Summary of scoring metric and source.

Metric	Unit	Video	Fluoroscopic Image
Tip-apex Distance (TAD)	mm		X
Duration of Procedure	minutes	X	X
Fluoroscopic Shots	count	X	X
Pullbacks	count	X	X
Head Cortex Breaches	count	X	X
Switch AP/Lateral Shot	count	X	X
Supervision Impact Score	formula	X	

The degree of intervention of the surgeon's supervisor was scored using a newly developed system for this purpose and applied based on tallies of the number of occurrences of each of three different behaviors: 1) supervisor instructions to someone other than the operating surgeon (e.g., the supervising surgeon requesting a fluoroscopic image from the radiology technician), 2) supervisor instructions to the operating surgeon (e.g., instructions to drill the wire deeper), and 3) the supervisor handling an instrument

as more than an helper to the surgeon (e.g. physically controlling the drill). Four raters independently tallied the number of supervisor interventions of each type for two representative cases, based on both the video recording and intraoperative fluoroscopic images. For each case a supervision intervention score was calculated by multiplying the tally of each intervention type with the weights shown in Table 2.ii and summing the result.

Table 2.ii Categorical weighting of supervision intervention.

Impact	Weight/Incidence
Instructions to others in room	1
Instructions/Tips	2
Handle of Instrument (more than assisting)	3

The TAD was measured from DICOM files of the final AP and lateral images of the wire navigation using OsiriX® software (Foundation OsiriX, Geneva, Switzerland). Ten raters independently measured the TAD on 7 cases using the protocol described in Johnson et al. (2008). An ellipse was fit to the subchondral bone of the femoral head and a line drawn between the two points where the ellipse met the femoral neck. This line was then bisected with a second perpendicular line, and where that line intersected the superior portion of the ellipse was used to define the apex.[36] Two of the graders repeated the measurements at least 3 weeks after the initial measurements.

Statistical Analysis

The repeatability of the TAD measurement was assessed with the standard deviation and inter-rater reliability of the 10 raters' TAD measurements for each case

(210 data points). Individual (AP and lateral) elements of the TAD were evaluated, as well as the total TAD (the sum of the AP and lateral elements). The intra-rater reliability of the TAD measurement was assessed with the two raters' first and second measurements (84 data points). The repeatability of the video metrics was assessed with the standard deviation and inter-rater reliability of the 4 raters' metrics. Intra-rater and inter-rater reliability analysis was assessed with both Cronbach's Alpha and Intraclass Correlation Coefficient (two-way random model with absolute agreement, 95% confidence). Reliability analysis was done using SPSS Software (version 23, IBM, Armonk, NY). Descriptive statistics were obtained using Minitab (version 17, Minitab, Inc., State College, PA).

2.3 Results

TAD results

For all graders the average standard deviation for AP, lateral, and summed TAD measurements were 2.7, 1.9, and 3.7 mm, respectively (Table 2.iii).

For the AP measure of case 1, there was one (out of 140) extreme outlying measurement (over 3.5 times the average and double the next highest value) that disproportionally skewed the standard deviation. With this single outlying value removed, the average AP standard deviation for case 1 was reduced from 6.0 to 2.5 mm and for all AP cases from 2.6 to 2.1 mm. The sum standard deviation for case 1 is reduced from 6.9 to 3.1 mm and the average for all cases from 3.7 to 3.1 mm.

Table 2.iii Mean and standard deviation in TAD measurements of 7 cases by 10 raters.

	AP	**Lateral**	**Sum**
Case	Mean (Standard Deviation) (mm)		
1	6.2 (6.0)	4.9 (1.2)	11.1 (6.9)
2	5.2 (1.5)	9.6 (1.8)	14.8 (2.3)
3	8.8 (3.8)	10.4 (3.4)	19.2 (5.6)
4	4.6 (1.3)	3.9 (1.8)	8.5 (2.5)
5	12.0 (1.6)	11.7 (2.6)	23.7 (2.8)
6	8.2 (2.8)	9.2 (0.9)	17.4 (3.3)
7	7.7 (1.9)	9.8 (1.6)	17.5 (2.3)
average	**7.5 (2.7)**	**8.5 (1.9)**	**16.0 (3.7)**

The inter-rater reliability analysis for all measures together (AP, lateral, and sum) and across the 10 raters produced a Cronbach's Alpha of 0.97 and Intra-class Correlation Coefficient of 0.72 for single measures and 0.96 for average measures. For intra-rater reliability between first and second measures of the same TAD with the two raters who repeated their measurements after two weeks, rater 1 had an average standard deviation of 1.39 mm and rater 2 had an average standard deviation of 1.92 mm. Rater 1's Cronbach's Alpha and Intra-class correlation coefficient (single measures and average measures) were 0.94, 0.89 and 0.94. Rater 2's Cronbach's Alpha and Intra-class correlation coefficient (single measures and average measures) were 0.88, 0.79, and 0.88.

Video Metric results

A summary of the metrics obtained from the 4 raters on 2 cases of the surgical video and intraoperative fluoroscopic images can be found in Table 2.iv. Time (from

fluoroscopic timestamps), number of shots, switches between AP and lateral views, and cortex head breaches were consistent for all raters on both cases. Time was also recorded by raters using the video timestamps (results not shown); this approach from estimating duration produced discrepancies between 2 to 21 seconds for case 1 and between 1 to 25 seconds for case 2. For both cases, the number of pullbacks, instructions to others in room, instructions, and supervising surgeon impact score varied between raters. The average standard deviation for instructions to others, instructions to the operating surgeon, handling of instruments and impact score for both cases was 1.64. For case 1 the standard deviation of the number of pullbacks was 0.5 for case 1 and 1.7 for case 2.

The inter-rater reliability analysis for all 9 metrics across both cases produced a Cronbach's Alpha of 0.99 and Intraclass Correlation Coefficient of 0.90 for single measures and 0.99 for average measures. The inter-rater reliability analysis for pullbacks alone for both cases produced a Cronbach's Alpha of 0.33 and Intraclass Correlation Coefficient of 0.95 for single measures and 0.30 for average measures. The inter-rater reliability analysis for the supervising surgeon impact score alone for both cases produced a Cronbach's Alpha of 0.99.

Table 2.iv Summary of video metrics, categories with discrepancies are in bold.

Case	Rater	Time (min)	Shots	Switch AP/Lateral	Breach cortex	**Pull backs**	**Instruction to others**	**Instruction**	**Handle Instruments**	**Score**
1	1	8.6	22	1	0	1	1	3	0	7
	2	8.6	22	1	0	2	1	3	0	7
	3	8.6	22	1	0	2	0	3	0	6
	4	8.6	22	1	0	2	0	1	0	2
			mean (std. deviation)			1.8 (0.5)	0.5 (0.6)	2.5 (1.0)	0.0 (0.0)	5.5 (2.4)
2	1	17.1	27	4	0	2	10	15	0	40
	2	17.1	27	4	0	1	9	17	0	43
	3	17.1	27	4	0	3	5	13	2	37
	4	17.1	27	4	0	5	6	18	1	45
			mean (std. deviation)			2.8 (1.7)	7.5 (2.4)	15.8 (2.2)	0.8 (1.0)	41.3 (3.5)

2.4 Discussion

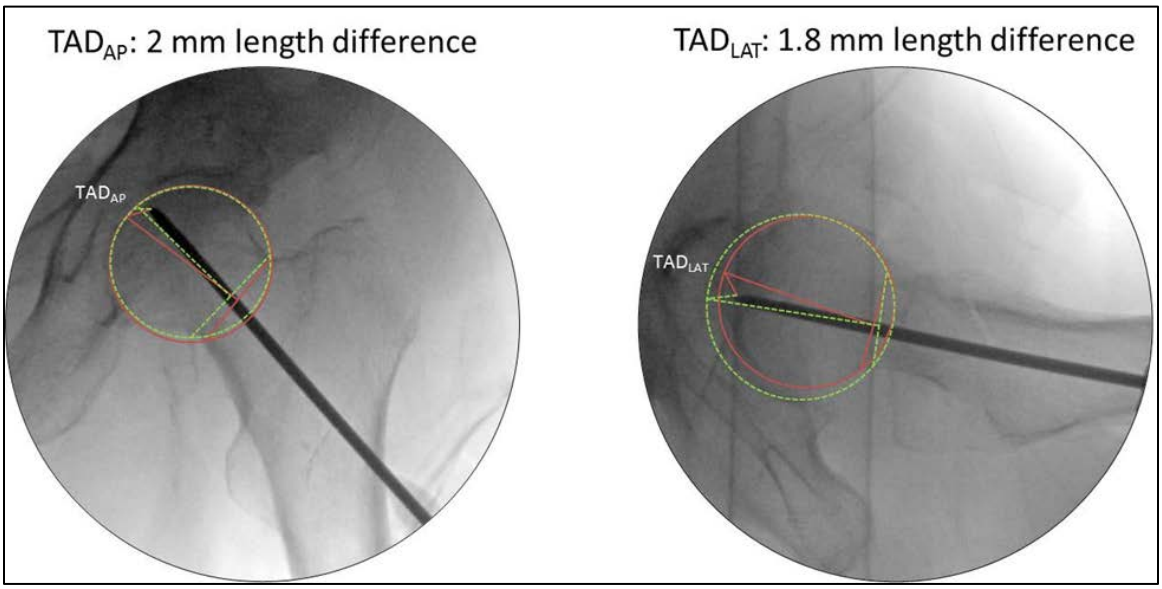
The high values of inter-rater reliability (0.97 Cronbach's Alpha) and intra-rater reliability (0.88 and 0.94 Cronbach's Alpha) show that the technique for measuring the TAD from intraoperative fluoroscopic images is valid and reproducible. The magnitude of the TAD sum standard deviation (3.1 mm) for 10 raters across 7 cases seems acceptable given the utility of the measure and keeping in mind that the error is relatively small compared to the criteria of 25 mm typically applied for clinical acceptance.

The video metrics exhibited high reproducibility between independent raters. Both cases had the same values across all raters for 4 of the 9 metrics. The number of pullbacks was less consistent than some of the other measures, with inter-rater reliability Cronbach Alpha of 0.33. We conclude that the number of pullbacks is unreliable and should not be used as an assessment metric. The lack of instances where the guide wire breached the cortex of the femoral head limits the study of that parameter, since no occurrences were identified from video or intraoperative images for the cases studied. Both pullbacks and cortex head breaches rely on determining a wire's position, which proved to be difficult: the drill tends to obscure the line of sight in the video images and the fluoroscope images are not continuous, so a head breach or repositioning may not be explicitly captured. Objectively determining whether a wire was removed and re-inserted into a previous hole in the cortex or versus starting a new cortex breach is difficult as the C-arm position may change between images. The timestamps from the fluoroscopic images proved to be a more convenient and consistent measure of surgery duration than measuring duration from the video.

Although tallies for each individual supervising surgeon categories varied between raters, the supervising impact score, which is the weighted sum of the three categories, exhibited a Cronbach Alpha inter-rater reliability measure of 0.99, suggesting a potentially useful new measure of surgeon supervision. Baumgaertner et al.'s seminal TAD study measured the inter-rater variability with the average TAD standard deviation. Their reported value of 1.7 mm is much smaller than the standard deviation of 3.1 mm observed in this study. The greater variability in the present study may be attributed to differences between placing K-wires and placing the implant. The variability may also be attributed to using more raters (ten rather than two), or because the raters were less experienced with the measure than those in the Baumgaertner study. Since each rater viewed the same images, the variability is a consequence of differences in matching an ellipse to the femoral head subchondral bone and/or intersection of this ellipse with the femoral neck (Figure 2.iii).

Baumgaertner et al. measured the intra-observer with the standard deviation of the averaged sum measure at two different time points. Again their report of 1.2 mm is smaller than the 2.1 mm observed in this study. The difference was clinically relevant: in one of the seven cases, the case-specific inter-observer standard deviation of 2.8 mm was large enough to yield erroneous conclusions on the mean sum measure of 23.7 mm when compared to the 25 mm threshold.

Figure 2.iii Example variation in AP (left) and lateral (right) TAD measurements due to differences in the geometric construction.



Using edited video of 25-40 minutes in length, Birkmeyer et al. showed that surgical skill assessed from video recordings of operations correlates to complication rate. Their sensitivity analysis showed that the video analysis had low rater bias and showed that repeated ratings of the same surgeon in different videos were highly correlated, with mean skills rating correlation of R=0.94.[21] The Birkmeyer study also demonstrated that lower technical skill is associated with a 40% longer operating time.[21] The metrics selected in this study are different than those used in the Birkmeyer study because they focus on orthopedic surgery rather than bariatric surgery, but they show a similar robustness and repeatability. Future studies will reveal whether they have the same consistency across physicians and whether they correlate with other measures of operating room performance.

OSATS, the gold standard measure of surgical performance, has been extensively studied; OR video assessment using OSATS global rating scale has an inter-rater reliability ranging from 0.57 to 0.83.[1] Video review utilizing OSATS interval assessment of resident performance had inter-rater reliability of 91±4.[37] The Ottawa Score, which uses a 5-point scale assessing readiness for independent practice, had a standard deviation for technical performance score of 1.01.[38] In laparoscopic surgery a 6-point scale with the amount of support required had an inter-rater reliability score of 0.87. The inter-rater Cronbach's Alpha for the 4 raters on the 2 cases for all video metrics (time, number of shots, pullbacks, cortex head breaches, switch between AP/lateral shots, supervisor instruction to others, supervisor instructions to surgeon, supervisor handle of instruments, and overall supervisor impact score) was 0.99, which compares quite favorably to these other techniques, suggesting that the techniques proposed may enjoy even greater inter-rater consistency than OSATS.

Limitations

The TAD measurements previously reported were made for the cannulated lag screw that is placed over the guide wire that we used for our measurements. As the guide wire is used to place the implant, we believe the measure provides an accurate representation of the implant position. Lag screws are available in increments of 5 mm, which limits the surgeons' accuracy in achieving an ideal TAD compared to the guide wire. The TAD measure is sensitive to details of the fluoroscope position and the quality of the radiographic image. If the resident alters the wire position after the final AP and lateral shots, this change in position is not accounted for with the TAD measure. The

present study relies on a limited number of cases, all from a single hospital. In some cases, videos were occasionally obscured or did not capture all areas of interest, and audio cues were occasionally unclear, which limited uniform measurement. We made no attempt to account for differences among the teaching styles (i.e., proclivity toward intervention) of the supervising surgeons, but instead attempted to measure the degree to which such interventions occurred in any given surgery. Resident and supervising surgeon behaviors may be altered due to the presence of the camera in the OR, although neither was aware of the grading protocol at the time of recording.

2.5 Conclusions

This study included four raters viewing the fluoroscopic images and video collected from a head-mounted camera for seven surgeries. The raters independently rated nine metrics from the data. Four of these, the surgery duration, the number of fluoroscope images collected, the number of C-arm position changes, and the degree of surgeon intervention (a weighted sum of three categories of intervention), were consistent across the four video reviewers and are likely to be useful for performance assessment. The tip apex distance was less reliable than previous reports have suggested, but is still a valuable metric of surgeon skill. Our study shows that video recording assessment allows non-experts to reliably measure these metrics, they offer an opportunity for objective, consistent assessment of operating room performance.

Chapter 3: Metric Correlation to Experience – Measures of hip fracture wire navigation performance in the operating room reflect surgical experience

Leah K. Taylor MS, Geb W. Thomas PhD, Matthew D. Karam MD, Clarence D. Kreiter PhD, Donald D. Anderson PhD

3.1 Introduction

Orthopedics has lagged behind other surgical disciplines in the development and adoption of simulation approaches to skills training and assessment. The American Board of Orthopedic Surgery requires applicants for certification to have successfully completed an accredited residency in which their surgical skills have been appropriately certified by a Program Director. The American Board of Surgery additionally requires applicants for certification in general surgery to have satisfactorily completed a Fundamentals of Laparoscopic Surgery program, which relies upon a validated laparoscopic trainer to simulate and assess performance.

Evidence for the benefits of simulation in training technical skills is accumulating, and lab based training is quickly finding its way into orthopedic residency programs [39; 40]. However, surgery performed in the operating room (OR) on a live patient is a unique experience, and there is no substitute for assessing performance in that environment. Furthermore, establishing skill transfer from a simulated setting to the OR is paramount in the development of simulator-based competency assessments [5]. Determining an unbiased, repeatable process for assessing surgical performance in the OR is a critical next step toward that goal.

Video recordings of surgery provide a platform for assessing surgical performance in the OR [21; 41; 42], but the community needs to decide how to assess OR performance from those recordings and what surgeries are best to record. The Objective Structured Assessment of Technical Skills (OSATS) is one widely used method for assessing surgical skill. A global rating scale is used to structure expert evaluation of task performance working from a list of operative competencies that are each rated on a 5-point Likert scale, anchored by behavioral descriptors. Recent studies have highlighted shortcomings in the objectivity of OSATS evaluations and in its ability to evaluate surgical outcome [1; 5; 16; 17]. Improved methods are needed to assess skills competency in the setting of direct patient care, especially considering that 63.5% of surgical errors are technical in nature [43].

The surgical treatment of hip fractures offers a promising target for OR performance assessment because it is common [21] and involves unbiased performance metrics that may be reliably quantified from video recordings and intraoperative fluoroscopic images [24]. The ability to navigate a wire within bone using fluoroscopy (wire navigation) is a challenging, generalizable skill and a critical step in fixation of these fractures with devices such as the Dynamic Hip Screw (DHS) or Telescoping Screw Plate (TSP). The success of a DHS surgery depends upon mechanically beneficial placement of the implant, reflected by the tip-apex distance (TAD) that estimates the distance from the tip of the implant to the apex of the femoral head [25]. A TAD $\leq$ 25 mm significantly decreases the risk of implant cut-out from the bone [26]. Adept OR performance of wire navigation involves balancing precision of wire placement with

procedure duration and the amount of fluoroscopy used. The duration correlates with both cost and patient outcomes [28], and adverse effects of radiation exposure are evident [29-31]. The present study addresses the following question: Do measures of hip fracture wire navigation performance, assessed from intraoperative video and fluoroscopic images, correlate with surgical experience?

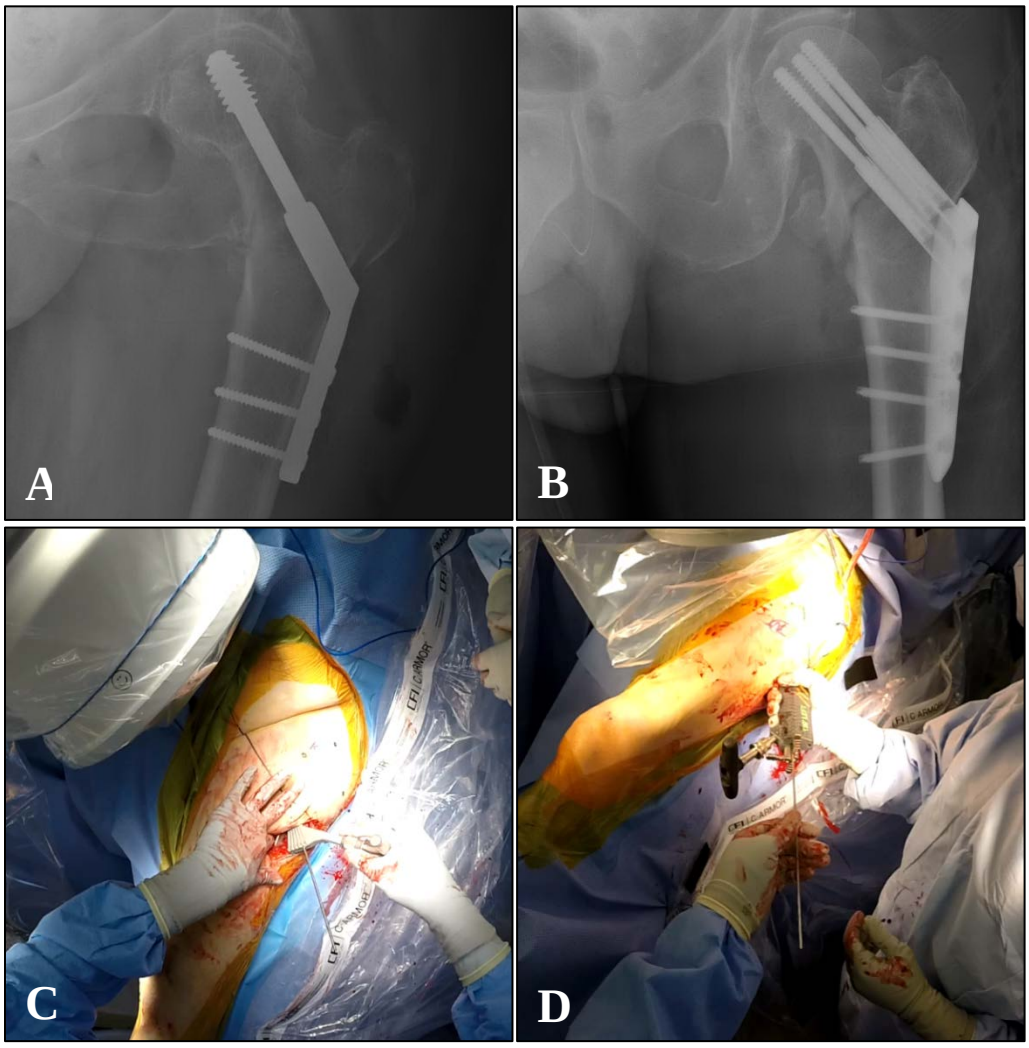
3.2 Methods

This prospective study included hip fractures (intertrochanteric and femoral neck) scheduled for open reduction and internal fixation using a DHS or TSP at the University of Iowa Hospitals and Clinics between August 2014 and March 2016. The DHS construct uses a single large screw (Figure 3.i-A), and the TSP three smaller parallel screws (Figure 3.i-B), to stabilize the hip fracture. These devices were chosen because a critical early step in both procedures is the placing of a guide wire in the proximal femur; a task completed using the assistance of a drill guide and fluoroscopic visualization (Figure 3.i-C&D). The IRB at the University of Iowa approved the protocol, and informed consent was obtained from patients and surgeons.

Data were collected for 18 cases, but three were subsequently excluded from analysis; one because it involved a very unstable fracture that confounded wire placement, another because the supervising surgeon was emergently called away during the wire navigation, and a third because the supervising surgeon was using the TSP for the first time, which substantially increased the duration of the surgery and added discussion time related to the unfamiliar implant. One staff surgeon and 13 different residents functioned as the operating surgeons (1 resident completed 2 procedures, both

as a third-year resident). Three residents were female, 10 were male, and the staff surgeon was male. The number of weeks into residency and the number of DHS and TSP cases previously logged by each surgeon were used as indicators of surgical experience. The number of cases came from the resident's ACGME case log. The staff surgeon's logged cases number (24) was taken from a billing log felt to most appropriately represent his considerable prior case experience.

Figure 3.i Postoperative radiographs of the two implants: (A) DHS and (B) TSP. Intraoperative images showing the guide with guide wire placed on the lateral cortex of the femur for the (C) DHS and (D) TSP.



Hip fracture cases were recorded in the OR using a GoPro® Hero3+Silver Edition (GoPro; San Mateo, CA) head-mounted point-of-view camera, and all intraoperative fluoroscopic images were automatically saved. A team member viewed and cropped the video to include only the wire navigation portion of the surgery. Data collected from the video were: duration of wire navigation (time from first fluoroscopic image with the guide wire to the final image of the wire ready for implant), number of fluoroscopic images collected, and the degree of intervention by the surgeon's supervisor.

The degree of supervisory intervention metric was created as an indicator of resident skill and readiness for independent practice, but it is also unavoidably sensitive to the interaction style of each supervisor. Any OR performance assessment involving residents must account for intervention by the supervising surgeon that influences the progression of the procedure. The degree of intervention of the resident's supervisor was measured using the sum of weighted tallies of three behaviors: weight 1 - supervisor instructions to someone other than the operating surgeon (e.g. requesting a fluoroscopic image from the radiology technician), weight 2 - supervisor instructions to the operating surgeon (e.g. instructions to alter the wire trajectory), and weight 3 - the supervisor handling an instrument in a capacity exceeding that of an assistant (e.g. physically controlling the wire guide: Figure 3.ii). For trend analysis purposes, the supervisors were grouped into two categories; a senior level resident or a staff surgeon.

The TAD was measured on the final anterior-posterior and lateral fluoroscopic images of the wire placement using OsiriX® software (OsiriX Foundation; Geneva, Switzerland) [36]. In a prior investigation of the reliability of our measurement methods, we

found that the inter-rater reliability of the TAD was 0.97 and of the video metrics (duration, number of images, and supervisory intervention score) was 0.99 [10]. The TAD measure was excluded for three cases where the supervising surgeon took control of the drill to place the guide wire in the final position. A supervisory intervention score was not assigned for the one case in which a staff surgeon acted as the operating surgeon.

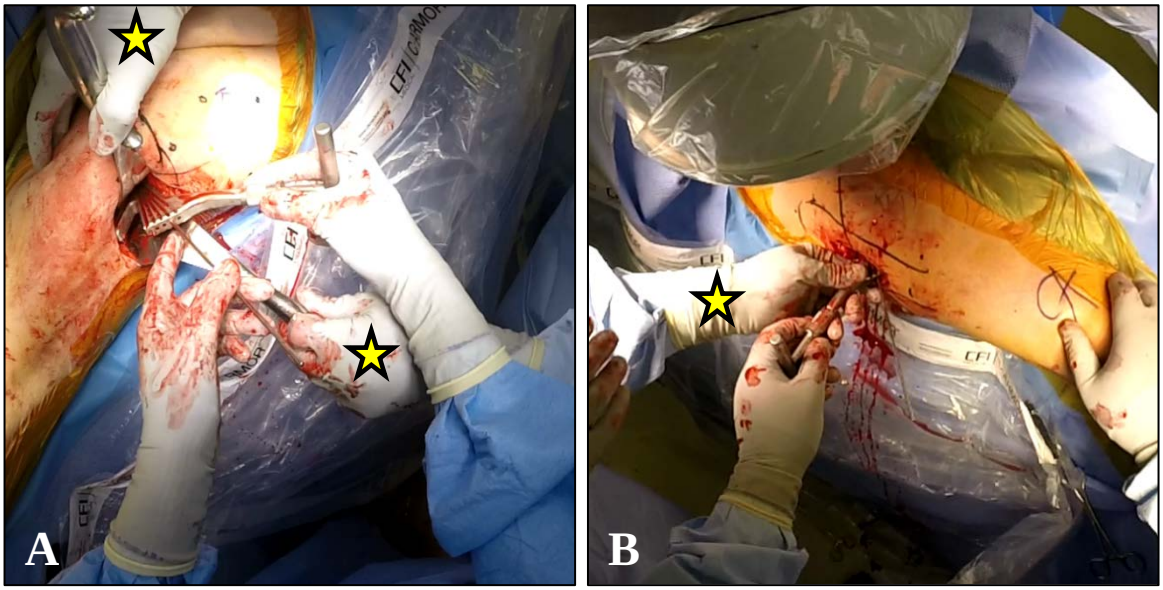
To relate the objective metrics of performance assessment to expert perception of skill, two fellowship-trained traumatologists independently graded each video using a modified OSATS scoring system [44]. Intraoperative fluoroscopic images were inset into the videos at the point in the surgery where obtained, and the videos were viewed on a secure website. The OSATS scores reflected only the wire navigation portion of the surgery. The average of the total OSATS score (maximum 45 points) from both raters was used for each case.

Statistical Analysis

Duration, number of fluoroscopic images obtained, supervision intervention score, and TAD were each correlated with the experience metrics (weeks in residency and the number of previously logged cases). As data were not normally distributed, Spearman rank correlations were used to assess all relationships. For measures found to have significant correlation, the Mann-Whitney U test compared differences in novice vs. expert performance: novice (PGY 2 and 3, n=10) and expert (PGY 5 and staff surgeon, n=5). Cronbach's α was used to assess interrater-reliability of the OSATS grading. Statistical significance was defined as $p<0.05$. Statistical analysis was completed with SPSS Software (version 23, IBM, Armonk, NY). A composite score was computed by

summing the average standardized values of the four performance metrics using SAS (version 9, SAS Institute, Cary, NC). Correlations between the composite score and experience were calculated using the same approach as for the individual performance metrics.

Figure 3.ii Examples of two supervision intervention behaviors. (A) The supervisor (hands indicated with stars) is handling the Bennett retractors in an assisting manner (elevation and retraction), and no supervision impact is counted. (B) The supervisor (on the left, hand indicated with a star) is taking control of the guide, which constitutes handling an instrument and is tallied with a weight of 3.



3.3 Results

Patient ages ranged from 55 to 97 years with a mean (standard deviation) of 79.3 (11.5) years. Patient BMI was 25.6 (6.9), and 67% of the patients were female.

Table 3.i presents the experience and performance metrics for the analyzed cases. The number of previous cases logged for residents ranged from 1 to 13 with an average of 5.7 (4.1) cases. Table 3.ii presents the correlation coefficients relating performance to experience metrics. Wire navigation duration was significantly correlated with both weeks into residency -0.66 ($p<0.01$) and prior cases logged -0.59 ($p=0.02$) (Figure 3.iii). The number of fluoroscopic images and the supervision intervention score did not correlate with either experience metric. TAD was significantly correlated with cases logged -0.67 ($p =0.02$) but not weeks into residency (Figure 3.iv). Mann-Whitney U analysis indicated that the wire navigation duration for the novice group was significantly higher than for the experts $U=9$ ($p=0.05$). There was no significant difference in the TAD between the two experience groupings. The composite performance metric significantly correlated to both weeks into residency -0.55 ($p=0.03$) and cases logged -0.66 ($p=0.01$) (Figure 3.v).

The inter-rater reliability of the two traumatologists' OSATS total scores was 0.71. There was no significant correlation between the OSATS total score and experience metrics (correlation coefficients: weeks into residency 0.43, $p=0.11$ and cases logged 0.43, $p=0.11$). Total OSATS score significantly correlated with duration -0.52 ($p=0.05$) and number of fluoroscopic images -0.83 ($p<0.001$), but correlated neither with supervision intervention score -0.27 ($p=0.36$) nor TAD -0.30 ($p=0.34$).

Table 3.i. Experience and performance metrics for 15 hip fracture cases listed in chronological order. Case 8 was completed by a staff surgeon (therefore no supervisor or supervision intervention score is shown). For cases 1, 4, and 10 the supervising surgeon placed the final guide pin. Cases 10 and 13 were completed by the same resident.

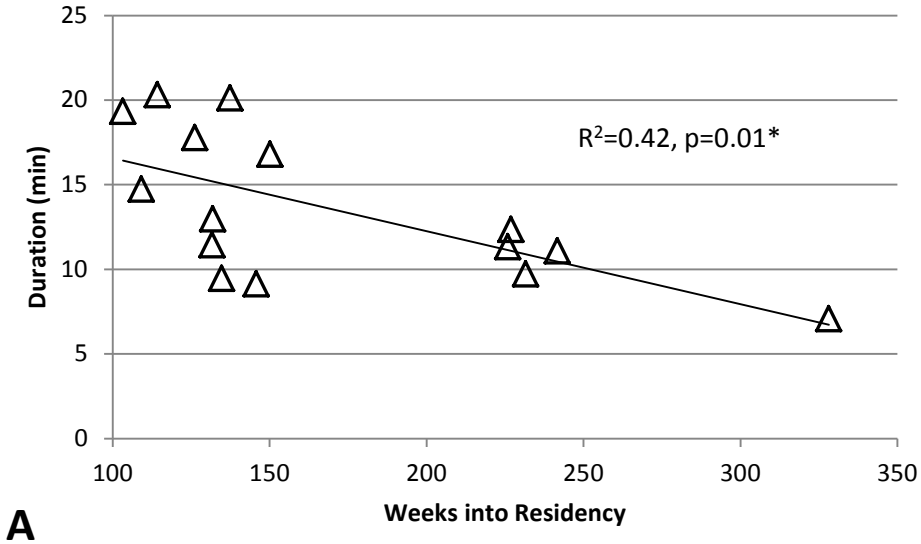
			Experience Metrics		**Performance Metrics**				
Case	Implant	Supervisor	Weeks into residency	#DHS Logged	Duration (min)	Fluoro Images	Supervision Intervention	TAD (mm)	Composite Performance
1	DHS	Resident	109	5	14.7	40	107	*	0.311
2	DHS	Resident	114	3	20.3	44	54	17	0.379
3	DHS	Staff	227	10	12.4	58	7	17	-0.123
4	DHS	Resident	135	5	9.5	19	27	*	-1.083
5	DHS	Staff	146	9	9.1	25	6	9	-1.194
6	DHS	Staff	150	1	16.8	26	43	23	0.136
7	DHS	Resident	103	1	19.3	68	65	17	0.739
8	TSP	*	328	24	7.1	37	*	10	-0.987
9	TSP	Staff	226	12	11.3	32	45	12	-0.608
10	DHS	Resident	126	2	17.8	55	142	*	1.100
11	TSP	Staff	232	7	9.7	45	43	25	0.104
12	DHS	Resident	132	1	11.5	37	127	19	0.255
13	TSP	Staff	132	3	13.0	79	107	20	0.908
14	TSP	Staff	242	13	11.1	27	65	12	-0.587
15	TSP	Resident	137	8	20.1	41	137	11	0.483

* Indicates field not applicable to case

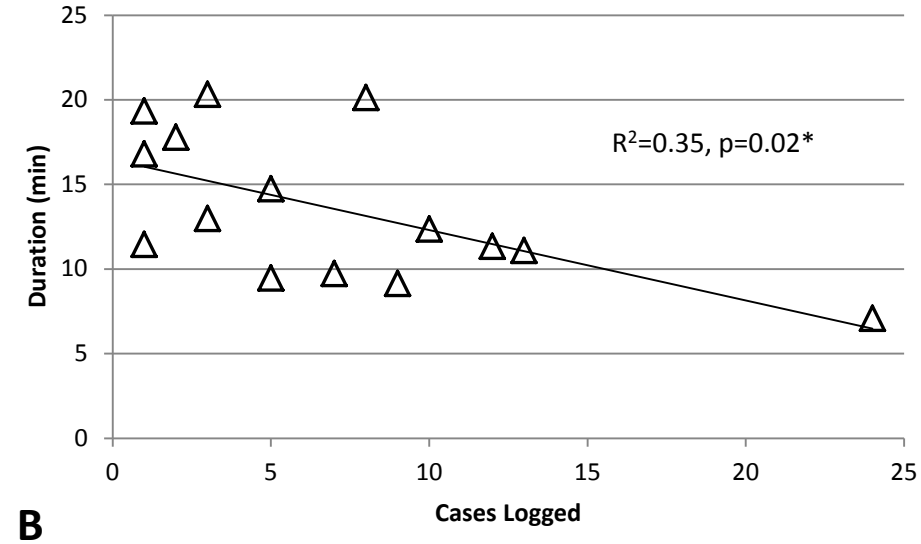
Table 3.ii Correlation coefficients relating performance metrics to surgeon experience. Significant correlations ($p<0.05$) are bolded.

		Experience Metrics	
Performance Metrics		Weeks into residency	Cases Logged
Duration (min)	Correlation Coefficient	**-.661**	**-.587**
	Significance, p-value	**0.007**	**0.021**
	n	**15**	**15**
Fluoroscopic Images	Correlation Coefficient	-0.34	-0.268
	Significance, p-value	0.216	0.335
	n	15	15
Supervision Intervention	Correlation Coefficient	-0.485	-0.346
	Significance, p-value	0.079	0.226
	n	14	14
TAD (mm)	Correlation Coefficient	-0.201	**-.669**
	Significance, p-value	0.531	**0.017**
	n	12	**12**
Composite Performance Metric	Correlation Coefficient	**-0.549**	**-0.656**
	Significance, p-value	**0.034**	**0.008**
	n	**15**	**15**
OSATS	Correlation Coefficient	0.044	0.092
	Significance, p-value	0.886	0.765
	n	13	13

Figure 3.iii Plots of *duration* vs. (A) weeks into residency and (B) number of previous cases logged. Results of linear regression are shown as a general indicator of the relationship.

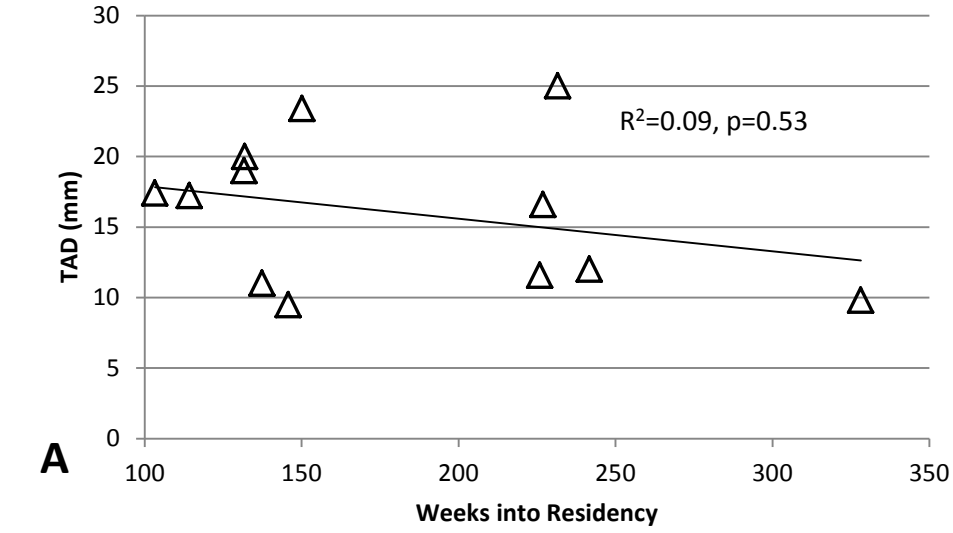


A

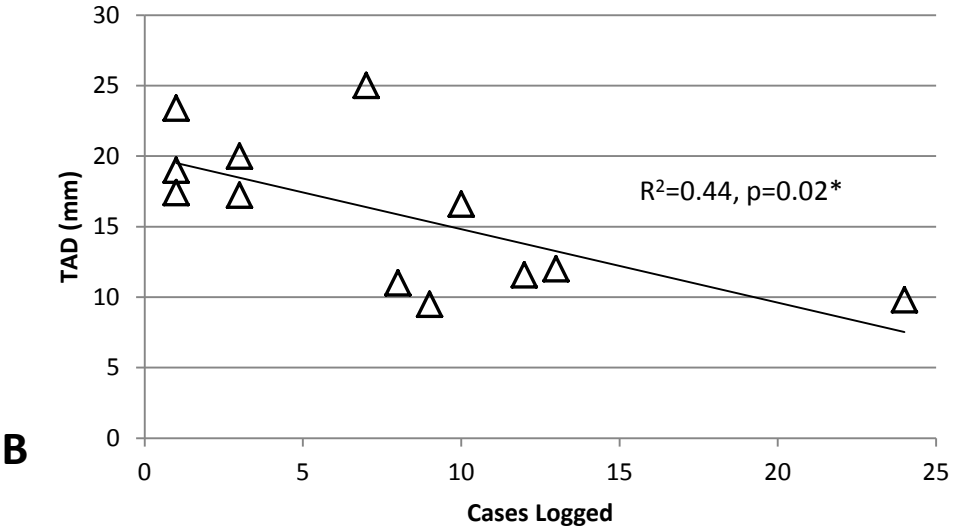


B

Figure 3.iv Plots of *TAD* vs. (A) weeks into residency and (B) number of previous cases logged. Results of linear regression are shown as a general indicator of the relationship.

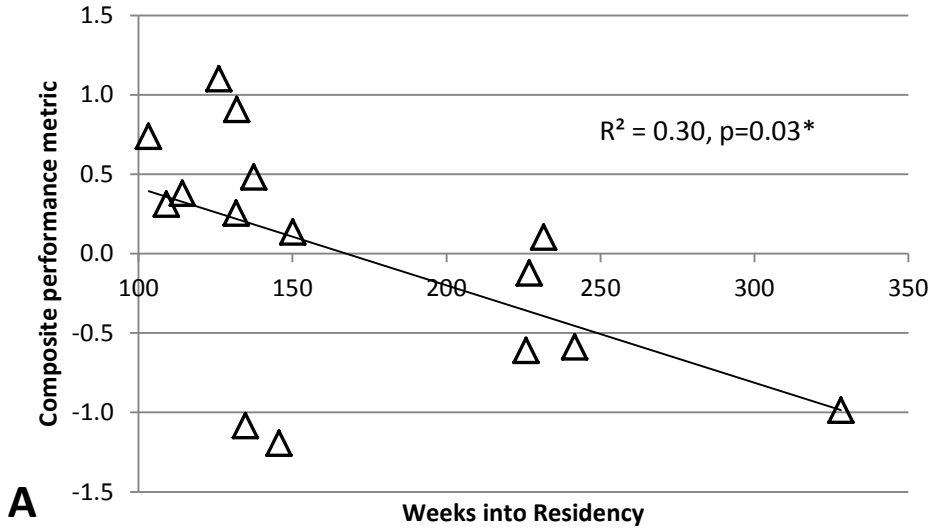


A

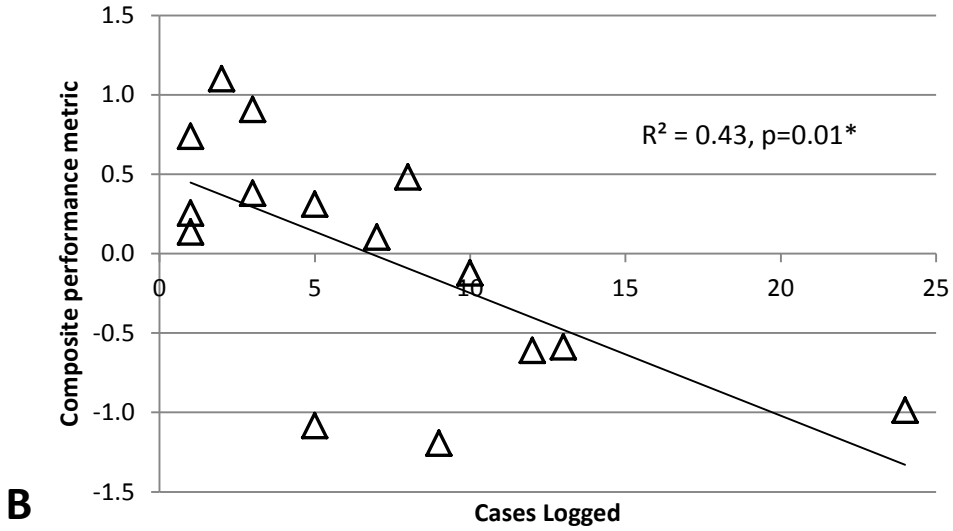


B

Figure 3.v Plots of the *composite performance metric* vs. (A) weeks into residency and (B) number of previous cases logged. Results of linear regression are shown as a general indicator of the relationship.



A



B

3.4 Discussion

The results from the evaluation of video recordings of 15 hip fracture cases performed by staff and residents indicate that two individual metrics of hip fracture wire navigation performance, duration and TAD, significantly differentiate surgical experience. A composite score integrating multiple individual performance metrics also provided strong correlations with surgical experience. Duration of the wire navigation had the strongest correlation to surgical experience, yielding significant negative correlation to both experience metrics. This is consistent with previous studies finding that greater surgical experience was associated with shorter operations [21; 45]. The TAD also exhibited a significant negative correlation with the number of previous cases logged. This finding agrees with that of a simulated setting in which experienced surgeons obtained better wire placement than novices [24].

The stronger correlation seen between performance and the number of previous cases logged than to the point in residency is consistent with previous findings that surgical skill is not directly related to duration of surgical practice, but more to procedural volume. For relatively complex procedures, such as wire navigation, morbidity and mortality have been shown to correlate with surgical volume [21]. The lack of TAD correlation to point in residency could be due an uneven distribution of experience levels (mostly PGY3 residents). It should also be noted that all residents performed wire navigation with a TAD $\leq$ 25 mm, suggesting residents are meeting recommended placement accuracy.

There was a modest trend for more experience to correlate with fewer fluoroscopic images taken during the wire navigation phase of the surgery, consistent with prior work that showed an increase in surgical experience results in a decrease of radiation exposure [29; 33-35]. The lack of statistical significance in our findings could be due to the high variability in the number of images used by the junior level residents (PGY 2 and 3); the number ranged from 19 to 79. We would expect to see a stronger correlation with an increase in sample size, specifically adding to the number of senior level cases.

The supervision intervention score did not significantly correlate with any of the performance metrics, but negatively correlated with weeks into residency -0.49 ($p=0.08$). This negative trend supports the hypothesis that as surgeon experience increases, the supervision intervention score decreases, most likely because the resident is more prepared for independent practice. Interestingly, although an increase in supervision intervention score did significantly predict increased duration, there was no corresponding improvement in the TAD.

Other scales used to assess supervisory intervention on surgical performance, such as the O-SCORE, rely upon the supervising surgeon to grade his/her intervention [38; 46]. Such an approach may provide insight regarding the supervisor's role in helping residents to gain technical proficiency, but directly involving the supervisor limits objectivity. We adopted a novel method of objectively quantifying involvement of the supervising surgeon from the point of view of an outsider, which shows potential as a discriminating factor of surgical skill.

The composite performance score correlated with weeks into residency more highly than did any individual performance metric except duration. The composite score correlated with previous cases logged with significance greater than any of the individual performance metrics. These finding suggest that the composite metric is the most reliable correlate of experience and may be the most attractive target for establishing competency standards.

Although there were trends toward correlations between OSATS score and experience metrics, no correlations were significant. The total OSATS score correlated significantly with duration and number of fluoroscopic images, but not to TAD. As the clinical importance of TAD is well established, the lack of the ability of OSATS to distinguish wire placement performance is concerning. This failure of OSATS to correlate with a surgical outcome measure is consistent with work that has shown that OSATS fails to correlate to radiographic assessment of reduction [5] and quality of fixation [17]. The inability of OSATS to accurately measure the key surgical outcome (TAD) suggests that new measures of performance, such as those introduced in this study, are needed to capture skills that affect outcomes.

The main limitation of this study is that a relatively small number of cases were analyzed, all from a single institution. We are currently working towards extending the use of this performance assessment approach to additional institutions. Variation in teaching habits between supervising surgeons (e.g., proclivity toward intervention) was not accounted for in our analysis because of the small sample size. Only one resident acted as the operating surgeon for multiple cases (two), one at 126 weeks into residency

and one at 132 weeks. Including cases for individual residents at multiple time points in residency could provide insight into individual learning curves. The wire navigation portion of hip fracture surgery is most often performed by a PGY 3 resident at the University of Iowa; therefore the range of experience level is limited. Inclusion of more experienced surgeons, specifically practicing orthopedic surgeons, would strengthen the study by more definitively establishing expert metric standards.

As the two implants included are not identical, differences in performance metric scores due to implant type were possible. However, we saw no indication of implant bias impacting performance. The two traumatologists who rated the videos with OSATS were among the attending surgeons on the cases and were familiar with the residents, so a truly blinded grading was not possible. As the purpose of including the OSATS was merely to compare the presented metrics with the gold standard of surgical skill evaluation, we did not feel a complete blinded assessment was essential. The value of the experience metric of previously logged cases is dependent upon diligence of the residents reporting case involvement. No modifications were made for the variance in fracture severity and degree of displacement of the cases; therefore performance metrics did not account for case difficulty.

Our results indicate that two individual metrics of hip fracture wire navigation performance, duration and TAD, significantly differentiate surgical experience. A composite score incorporating multiple performance metrics also provided strong correlations with surgical experience. The methods presented have the potential to

provide truly objective assessment of resident technical performance in the OR, a critical step towards competency based certification.

Chapter 4: Conclusion

Our first hypothesis that wire navigation performance can be objectively collected using video and intraoperative images by non-experts was shown. The wire navigation duration, the number of fluoroscopic images used, the degree of surgeon intervention, and the tip apex distance were consistent and are likely to be useful for skill assessment. The second hypothesis that those reliable performance metrics are associated with surgical experience is also supported. Significant correlation was found between the number of previous cases performed and both surgery duration and tip apex distance.

Surgical skills training is the focus of our collaborative research team including members from Biomedical Engineering, Orthopedics and Rehabilitation, Industrial Engineering, and Family Medicine departments. This research plays in integral role in relating skill training in a simulated setting to the actual operating room. Most of the metrics assessed here are used in a wire navigation simulator developed by our group. This allows the same metrics measured on simulator performance to be measured in the operating room. The clinical applicability is evident and direct measurement in a simulated setting are possible for the metrics of duration, amount of fluoroscopy used and TAD. The supervision impact score is distinct in that it tackles one of the disparities between a simulated and operating room environment. Where complete independent practice is possible in the lab, the nature of the OR prohibits this, proving to be a disconnect between the two. For the patient's safety, supervising surgeons oversee and step-in as necessary. Performance is altered due to this interference; the impact score

quantifies the contribution of the supervisor versus the operating resident. Competency has been labeled as the ability to perform the task without supervision. [6] This metric is a means of evaluating independent practice in the less experimentally controllable OR environment. The link to the actual OR setting is critical in the success and development of simulation based medical education and is some of the first research in orthopedics validating performance transfer to the operating room.

Additional research will also explore the relationship of simulator performance to OR performance and provide guidance to residency training programs. In the near future, case collection will spread to institutions in the Midwest Orthopaedic Surgical Skills (MOSS) consortium. The inclusion of this data, with a diversity of residents, supervisors, and teaching environments from multiple institutions, will provide a foundation for credentialing bodies to consider setting performance standards for residents.

The assessment of wire navigation is presented here, but the value the recordings add to the educational experience of orthopedic residents is far reaching. At the University of Iowa, the adoption of point of view cameras in the OR has been overwhelming positive by residents and staff surgeons. These recordings provide platforms for self, peer, and supervisor review which enhance the resident training experience. A survey sent out to orthopedics residents showed that almost ninety percent believed intraoperative video such as this would provide value to their education and ranked self-review as the favored method for learning. [47] Partially due to the influence and success of this work, the orthopedic department is working towards residents collecting a portfolio of milestone cases to demonstrate surgical competence.

Trial and error work on the most appropriate protocol and technology such as cost, battery life, resolution, weight, and size (Table 4.i) was essential to the feasibility of this research. As technology continues to advance, new devices will be adopted, but the overall utility of the video review in the high stakes environment will remain. With the changing world of surgical education, tools and applications such as this will play an essential role in shaping the new educational paradigm.

Table 4.i Comparison of devices relevant specifications, with the outlined box indicating the device utilized.

	GoPro Hero3+ Silver [48; 49]	GoPro Hero4 Session [48; 49]	GoPro Hero 4 Silver [48; 49]	GoPro Hero4 Black [48; 49]	Google Glass [50; 51]
List price	$300	$200	$400	$500	*$1,500 (no longer available)*
Recording time (30 fps, maximum video resolution, Wi-Fi off)	3:00	2:05	1:50	1:30	*1:00 - 1:30*
Video resolution	1080p max FPS 60	1080p max FPS 60	1080p max FPS 60	1080p max FPS 160	720p
Weight/with housing	74g/136g	74g/89g	83g/147g	88g/152g	36 g
Dimensions	41x59x30 mm	38x38x38 mm	41x59x30 mm	41x59x30 mm	glasses

*This table was adopted from "Value Added: The Case for Point-of-View Camera Use in Orthopedic Surgical Education" [47]

Limitations

The limitation of case size remains the largest hurdle for this type of research to overcome, but we believe with the inclusion of multiple institutions an adequate sample size is possible. This research also highlights the utility of non-experts as raters. The term non-expert is intended to specify that the raters were not expert orthopedic surgeons. Raters were familiar with the operation including steps, equipment, and the various operating personnel roles. Someone who is unfamiliar with the procedure would need to be briefed prior to providing dependable ratings. For many surgical operations, a vendor representative is present in the operating room to provide guidance on the implant device; this includes instruction on steps, equipment function, sizing options, and other product expertise. An article published in the Washington Post drew a lot of negative attention to the role of sales representatives in the OR. The role and consequential impact of the vendor representative input was not accounted for with this research as the surgical intervention was counted solely from the overseeing surgeon. The actual operations and decisions are made by the certified surgeon, but the representative is a resource intended to enhance patient safety.[52] Raters, for both the OSATS and performance metric analysis, were not blinded to the residents' identities. With the nature of investigative research this was not possible, but as additional institutions are added, a true blinded rating is feasible.

An additional limitation of this work is the possibility of error introduced into the TAD measure by the nature of fluoroscopy [53]. The 2D view of the 3D bone may be distorted due to both projective and image intensifier distortion. Projective distortion is inherent to all x-ray modalities. Image intensifier distortion is specific to fluoroscopy and

includes pincushion effects and S distortion. There is less distortion closer to the center of the image. The majority of the images used for measuring the TAD were towards the center of the field of view as surgeons instruct the radiology technician to center on the femoral head.

The reconstruction of the TAD is assuming that the two images (AP and lateral) are perfectly perpendicular, which is not probable as the C-arm is swung manually. As the scale of accuracy necessary for a TAD measure below the threshold of 25 mm, this error should not have a significant impact. Each image was calibrated to the known diameter of the guide wire visible in the image (either 3.0 or 3.2 mm). A sample analysis showed that a 3.2 mm diameter wire represented approximately 9.0 pixels. Despite these limitations, fluoroscopic images are universal used for clinical treatment.

References

1. Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, et al. 2010. Objective assessment of technical surgical skills. Br J Surg 97:972-987.
2. Marsh JL. 2015. AOA 2014-2015 presidential address: "tipping points" in surgical education. J Bone Joint Surg Am 97-A:5.
3. Sonnadara RR, Mui C, McQueen S, et al. 2014. Reflections on competency-based education and training for surgical residents. J Surg Educ 71:151-158.
4. Long DM. 2000. Competency-based residency training: the next advance in graduate medical education. Acad Med 75:1178-1183.
5. Mayne IP, Brydges R, Moktar J, et al. 2016. Development and assessment of a distal radial fracture model as a clinical teaching tool. J Bone Joint Surg Am 98:410-416.
6. Dwyer T, Wadey V, Archibald D, et al. 2015. Cognitive and psychomotor entrustable professional activities: can simulators help assess competency in trainees? Clin Orthop Relat Res 474:926-934.
7. Chang J, Banaszek DC, Gambrel J, et al. 2015. Global rating scales and motion analysis are valid proficiency metrics in virtual and benchtop knee athroscopy simulators. Clin Orthop Relat Res 474:956-964.
8. Brauer CA, Coca-Perraillon M, Cutler DM, et al. 2009. Incidence and mortality of hip fractures in the United States. JAMA 302:1573-1579.
9. Ensrud KE. 2013. Epidemiology of fracture risk with advancing age. J Gerontol A Biol Sci Med Sci 68:1236-1242.
10. Taylor L, Thomas GW, Karam MD, et al. 2016. Assessing wire navigation performance in the operating room. J Surg Educ 73(5).
11. Kim MJ, Williams RG, Boehler ML, et al. 2009. Refining the evaluation of operating room performance. J Surg Educ 66:352-356.
12. Jeray KJ, Frick SL. 2014. A survey of resident perspectives on surgical case minimums and the impact on milestones, graduation, credentialing, and preparation for practice. J Bone Joint Surg Am 96:e195.
13. Marsh JL. 2013. Should time spent in residency define the end of training? J Bone Joint Surg Am 95:1905-1905.
14. Aggarwal R, Grantcharov T, Moorthy K, et al. 2007. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. Ann Surg 245:992-999.
15. Niitsu H, Hirabayashi N, Yoshimitsu M, et al. 2013. Using the objective structurued assesment of technical skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. Surg Today 43:5.
16. Hopmans CJ, Hoed PTd, Laan Lvd, et al. 2014. Assesment of surgery residents' operative skillls in the operating theater using a modified objective structured assessment of technical skills (OSATS): a prospective multiocenter study. Surgery 156:11.

17. Anderson DD, Long S, Thomas GW, et al. 2016. Objective structured assessments of technical skills (OSATS) does not assess the quality of the surgical result effectively. Clin Orthop Relat Res 474:874-881.
18. Vassiliou MC, Dunkin BJ, Marks JM, et al. 2010. FLS and FES: comprehensive models of training and assessment. Surg Clin North Am 90:535-558.
19. Alvand A, Khan T, Al-Ali S, et al. 2012. Simple visual parameters for objective assessment of arthroscopic skill. J Bone Joint Surg Am 94:e97.
20. Moorthy K, Munz Y, Sarker SK, et al. 2003. Objective assessment of technical skills in surgery. Brit Med J 327:1032-1037.
21. Birkmeyer JD, Finks JF, O'Reilly A, et al. 2013. Surgical skill and complication rates after bariatric surgery. New Engl J Med 369:1434-1442.
22. Beard JD, Jolly BC, Newble DI, et al. 2005. Assessing the technical skills of surgical trainees. Br J Surg 92:778-782.
23. Salazar D, Schiff A, Mitchell E, et al. 2014. Variability in accreditation council for graduate medical education resident case log system practices among orthopaedic surgery residents. J Bone Joint Surg Am 96:e22.
24. Thomas GW, Johns BD, Kho JY, et al. 2015. The validity and reliability of a hybrid reality simulator for wire navigation in orthopedic surgery. IEEE Trans Human-Mach Syst 45:119-125.
25. Baumgaertner MR, Curtin SL, Lindskog DM, et al. 1995. The value of the tip-apex distance in predicting failure of fixation of peritrochanteric fractures of the hip. J Bone Joint Surg Am 77:1058-1064.
26. Kumar AJ, Parmar VN, Kolpattil S, et al. 2007. Significance of hip rotation on measurement of 'tip apex distance' during fixation of extracapsular proximal femoral fractures. Injury 38:792-796.
27. Farnworth LR, Lemay DE, Wooldridge T, et al. 2001. A comparison of operative times in arthroscopic ACL reconstruction between orthopaedic faculty and residents: the financial impact of orthopaedic surgical training in the operating room. Iowa Orthop J 21:31-35.
28. Edwards C, Counsell A, Boulton C, et al. 2008. Early infection after hip fracture surgery: risk factors, costs, and outcome. J Bone Joint Surg Br 90-B:770-777.
29. Giordano BD, Grauer JN, Miller CP, et al. 2011. Radiation exposure issues in orthopaedics. J Bone Joint Surg Br 93:e69.
30. Maxon HR, Thomas SR, Saenger EL, et al. 1977. Ionizing irradiation and the induction of clinically significant disease in the human thyroid gland. Am J Med 63:967-978.
31. Hynes DE, Conere T, Mee MB, et al. 1992. Ionising radiation and the orthopaedic surgeon. J Bone Joint Surg Br 74:332-334.
32. Khan IA, Kamalasekaran S, Fazal MA. 2012. Risk of ionising radiation to trainee orthopaedic surgeons. Acta Orthop Belg 78:106-110.
33. Palácio EP, Ribeiro AA, Gavassi BM, et al. 2014. Exposure of the surgical team to ionizing radiation during orthopedic surgical procedures. Rev Bras Orthop 49:227-232.

34. Bar-On E, Weigl DM, Becker T, et al. 2010. Intraoperative C-arm radiation affecting factors and reduction by an intervention program. J Pediatr Orthop 30:320-323.
35. Blattert TR, Fill UA, Kunz E, et al. 2004. Skill dependence of radiation exposure for the orthopaedic surgeon during interlocking nailing of long-bone shaft fractures: a clinical study. Arch Orthop Trauma Surg 124:659-664.
36. Johnson LJ, Cope MR, Shahrokhi S, et al. 2008. Measuring tip–apex distance using a picture archiving and communication system (PACS). Injury 39:786-790.
37. Seymour NE, Gallagher AG, Roman SA, et al. 2002. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. Ann Surg 236:458-463; discussion 463-454.
38. Gofton WT, Dudek NL, Wood TJ, et al. 2012. The ottawa surgical competency operating room evaluation (O-SCORE): a tool to assess surgical competence. Acad Med 87:1401-1407.
39. Karam MD, Kho JY, Yehyawi TM, et al. 2012. Application of surgical skill simulation training and assessment in orthopaedic trauma. Iowa Orthop J 32:76-82.
40. Karam MD, Westerlind B, Anderson DD, et al. 2013. Development of an orthopaedic surgical skills curriculum for post-graduate year one resident learners - the University of Iowa experience. Iowa Orthop J 33:178-184.
41. Beard JD, Jolly BC, Newble DI, et al. 2005. Assessing the technical skills of surgical traineesDELETE. Br J Surg 92:778-782.
42. Guerlain S, Turrentine B, Adams R, et al. 2004. Using video data for the analysis and training of medical personnel. Cogn Technol Work 6:131-138.
43. Fabri PJ, Zayas-Castro JL. 2008. Human error, not communication and systems, underlies surgical complications. Surgery 144:557-565.
44. Karam MD, Thomas GW, Koehler DM, et al. 2015. Surgical coaching from head-mounted video in the training of fluoroscopically guided articular fracture surgery. J Bone Joint Surg Am 97:1031-1039.
45. Leong JJH, Leff DR, Das A, et al. 2008. Validation of orthopaedic bench models for trauma surgery. J Bone Joint Surg Br 90-B:958-965.
46. Miskovic D, Wyles SM, Carter F, et al. 2011. Development, validation and implementation of a monitoring tool for training in laparoscopic colorectal surgery in the English National Training Program. Surg Endosc 25:1136-1142.
47. Karam MD, Thomas GW, Taylor L, et al. 2016. Value added: the case for point-of-view camera use in orthopedic surgical education. Submitted to the Iowa Orthop J.
48. GoPro. Support Articles Camera Battery-life.
49. Vegasaur. GoPro Camera Comparison: Hero4 vs. Hero3+.
50. Makhni EC, Jobin CM, Levine WN, et al. 2015. Using wearable technology to record surgical videos. Am J Orthop 44:163-166.
51. Wikpedia. Google Glass.
52. Hilzenrath DS. 2009. Medical sales rep work alongside doctors, even in operating rooms. The Washington Post, December 27, 2009 ed.

53. Nickoloff EL. 2011. AAPM/RSNA physics tutorial for residents: physics of flat-panel fluoroscopy systems. RadioGraphics 31:591-602.