

Title: Validation of the updated ArthroS Simulator: face and construct validity of a passive haptic virtual reality simulator with novel performance metrics

Authors: P Garfield Roberts, P Guyver, M J Baldwin, K Akhtar, A Alvand, A J Price, J L Rees

Patrick Garfield Roberts – Corresponding Author

RCS Dinwoodie Simulation Research Fellow/DPhil Candidate

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

E-mail: patrick.garfieldroberts@ndorms.ox.ac.uk

Telephone: +44 (0) 79485 941 675

Paul Guyver

Surgeon Commander

MDHU Derriford, Plymouth Hospitals NHS Trust, UK

Matthew Baldwin

Orthopaedic Registrar

Health Education Thames Valley

Kash Akhtar

Senior Clinical Academic Lecturer

The Blizard Institute, Barts & The London School of Medicine and Dentistry, Queen Mary University of London, UK

Abtin Alvand

NIHR Clinical Lecturer

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

Andrew J Price

Professor of Orthopaedic Surgery

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

Jonathan L Rees

Professor of Orthopaedic Surgery and Musculoskeletal Science

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

Abstract

Purpose: To assess the construct and face validity of ArthroS, a passive haptic VR simulator. A secondary aim was to evaluate the novel performance metrics produced by this simulator.

Methods: Two groups of 30 participants, each divided into Novice, Intermediate or Expert based on arthroscopic experience, completed three separate tasks on either the knee or shoulder module of the simulator. Performance was recorded using 12 automatically-generated performance metrics, and video footage of the arthroscopic procedures. The videos were blindly assessed using a validated global rating scale (GRS). Participants completed a survey about the simulator's realism and training utility.

Results: This new simulator demonstrated construct validity of its tasks when evaluated against a GRS ($p \leq 0.003$ in all cases). Regards its automatically-generated performance metrics, established outputs such as time taken ($p \leq 0.001$) and instrument path length ($p \leq 0.007$) also demonstrated good construct validity. However, two-thirds of the proposed 'novel metrics' the simulator reports couldn't distinguish participants based on arthroscopic experience. Face validity assessment rated the simulator as a realistic and useful tool for trainees, but the passive haptic feedback (a key feature of this simulator) is rated as less realistic.

Conclusion: The ArthroS simulator has good task construct validity based on established objective outputs, but some of the novel performance metrics couldn't distinguish between surgical experience. The passive haptic feedback of the simulator also needs improvement. If simulators could offer automated and validated performance feedback, this would facilitate improvements in the delivery of training by allowing trainees to practice and self-assess.

Introduction

Present day orthopaedic trainees have less exposure to surgical cases over a briefer training period than their trainers did[25]. To compensate for these changes, training programmes are said to be more formalised and structured[18], and simulators are now being incorporated into orthopaedics as training, selection and assessment tools[24].

Virtual-reality (VR) simulation has focused on endoscopic procedures in other surgical disciplines[1, 31] and in orthopaedics this translates to arthroscopic constructs, which were some of the first VR simulators[39]. If simulation is to become incorporated into orthopaedic training, the need to develop an evidence-base for the use of different types of simulators is pressing. Simulators and their tasks should be *valid* representations of the environment they are simulating. Of particular interest during simulator development are *face* and *construct validity*. Face validity describes how realistic the simulator appears to the end user (usually trainees)[37] and affects their *acceptance* of the simulator as a useful learning tool. Construct validity describes how the simulator is able to distinguish participants with different levels of surgical expertise. These forms of validity and others[12] (see Table 1) ensure simulators are acceptable to trainees and assessors, useful for training, and ultimately result in improvements in real life surgical performance[8, 17].

In an attempt to improve VR simulation, the ArthroS VR knee and shoulder arthroscopic simulator (VirtaMed, Zurich, Switzerland) has undergone significant developments from the version of the simulator used in previous studies[13], responding to concerns raised about its face validity and tactile feedback[33]. Uniquely among commercially available simulators[2], it now uses adapted authentic instruments to improve fidelity, so trainees can familiarize themselves with appropriate equipment. The improvements including 'passive haptic' tactile feedback where the resistances encountered during the simulated procedure are the results of physical objects making contact with the surgical instruments, while the on-screen view remains a virtual reality image. Typically, previous haptic devices in virtual arthroscopy[2], laparoscopy[28] and neurosurgery[14] have given 'active' feedback (often called 'force feedback') where motors mimic the presence of physical objects by applying variable resisting forces to the instruments, and this is felt to be less realistic. The ArthroS simulator automatically reports performance metrics during the simulated tasks. While some of these metrics are recognized as established and valid outputs of arthroscopic skill[16], others are novel metrics, such as joint surface scratches. Some of these metrics are directly measured (e.g. time, path length), whilst others are mathematically derived and presented as scores out of 10. A similar combined VR/physical simulator concept has been developed in the PASSPORTv2 training environment which also evolved from a previous device[35]. It shows the benefit of introducing passive haptics at improving the face validity of VR devices[32], and also the benefit introducing novel metrics to simulators, providing performance feedback to the user on the pressure they are applying at the joint surface during the procedure.

This paper focuses on the novel features introduced in this version of the simulator not present in previous studies on the ArthroS Simulator[13, 33]. The primary aim was to assess this updated simulator (version 1.2) for construct and face validity of its passive haptic VR environment and define its acceptability to trainees as a training tool. The secondary aim was to assess the novel performance output metrics it routinely produces.

This study examines a new iteration of a VR simulator which has both novel and improved features. Its is important to ensure these features are educationally validated as a training tool to justify any investment in simulator technology, and to help us better understand how we can train and monitor out surgeons. This is especially true of the novel metrics, because if ‘high stakes’ decision making (e.g. clinician revalidation) is made on the basis of the results of these simulators, we must be confident as assessors they test and train the qualities we expect.

Methods and Materials

Institutional review board approval was granted for this study and all subjects gave informed consent to participate. The study took place in the orthopaedic surgical skills laboratory of a university teaching hospital. Sixty candidates were recruited and ranked into three groups based on arthroscopic experience (novice, intermediate or expert; see Table 2). These candidates were randomized to either the knee simulation tasks or the shoulder simulation tasks. No candidate participated in both. There were 15 in each Novice Group, 10 in each Intermediate Group and 5 in each Expert Group. The novice group contained medical students, foundation and core trainees with no arthroscopic experience. The intermediate group contained core surgical trainees and orthopaedic registrars who had performed up to 100 arthroscopies. The experts were senior fellows and consultants, who had performed over 100 arthroscopies. Participants were excluded if they had previous exposure to a VR simulator. The groups are deliberately selected to represent different levels of arthroscopic experience, with the expectation they will perform differently, and the understanding that is a simulation is ‘construct valid’ we can use the results to distinguish between the groups. As such, the groups participants are not the subjects of the test *per se*, but set the standard we use to interpret the validity of the simulation. If the results of the simulation are not different for three groups which are selected to represent different standards of performance, then the simulation is not ‘construct valid’ and not a useful tool for testing performance.

Simulator and tasks

The ArthroS virtual reality simulator with version 1.2 software (VirtaMed, Zurich, Switzerland) was used, with interchangeable external joint models for the knee and shoulder groups. Each model has pre-defined arthroscopy portals, typical arthroscopic instruments (camera, hook, shaver and grasper), and a high definition display to view the procedure. The simulator allows the user to perform a variety of training tasks using the knee and shoulder modules: 17 basic skills tasks, 18 diagnostic tasks and 18 therapeutic tasks, of varying difficulty. The three most relevant joint-specific tasks were selected from the knee and shoulder modules by expert consensus amongst the three most senior authors (JR, AP, KA). They were chosen to be the most representative of the UK ISCP curriculum outcomes for completion of orthopaedic training[6]. Guided tasks (where the simulator prompts the next step of the task) were chosen so the study could focus on the skill-based behaviour without confounding from the rule- and knowledge-based behaviours which would disproportionately affect novices. Each candidate received an instructional PowerPoint presentation introducing the simulator and demonstrating the tasks to perform, and then had 5 minutes to familiarise themselves with the simulator before the simulation commenced. A checklist of the steps required in each task was also displayed during the simulated procedure (Table 3).

Data Collection

For each participant, three sets of data were collected: the simulator produced a report of performance metrics for each task; an arthroscopic video feed of each task was recorded for blinded GRS assessment; and participants completed a survey after the last arthroscopic task.

Outcomes Measures

Simulator Reports

Upon completing each task, the simulator reports the time taken (minutes and seconds), and the path length travelled (centimetres, to 1 decimal place) by the tips of the arthroscopic camera and of any instruments used as calculated by the simulator using the distance travelled in VR space. Additionally, the simulator reports the area of joint surface scratched (calculated as a percentage of the total joint surface contacted by VR space collision

detection of the user controlled camera and instruments with the joint surface during the task) and the volume of torn or healthy meniscus debrided during the guided meniscectomy task (calculated as a percentage of the volume rendered in VR space). Finally, the simulator calculates ‘scores’ (out of 10) from each measured metric compared to a goal value using an inbuilt algorithm[36]. For the purposes of the results, time taken and path length are established metrics in the literature[11, 16], while joint surfaces scratches, volume of meniscus debrided and all of the calculated scores were considered novel metrics as these have not been validated.

Figure 1 shows an example of a simulator report following Guided Meniscectomy II Task on the Knee module, and how the various metrics are related.

Global rating scale

The recorded video output footage from the procedure was rated by two raters (KA and MB) blinded to the experience of the participants and trained in the use of the ‘arthroscopic surgical skill evaluation tool’ (ASSET). The ASSET is a global rating scale (GRS) validated to have high internal consistency and inter-rater reliability in simulated and operating room environments [21]. The blinded raters independently review the arthroscopic footage and rate the participants performance according to eight domains (Safety, Field of View, Camera Dexterity, Instrument Dexterity, Bi-manual Dexterity, Flow of Procedure, Quality of Procedure, Autonomy) and an additional control domain for ‘Added Complexity of Procedure’. Most domains are rated on a scale of 1 (lowest) – 5 (highest), with anchor statements corresponding to novice, competent and expert performance according to the Dreyfus model of skills acquisition at 1, 3 and 5, except Autonomy and Added Complexity or Procedure which are rated out of three. During knee and shoulder diagnostic tasks, the ASSET domain for Bimanual Dexterity was excluded. In all tasks, ASSET domains for Autonomy and Added Complexity were excluded, as no help was available to participants, and the tasks were identical between participants.

Face validity questionnaire

To assess face validity, each participant answered a 13-item questionnaire immediately upon completing the tasks. This survey used a 5-point Likert scale with anchor statements (1-Strongly Disagree to 5-Strongly Agree) to evaluate the realism of the simulator and its training utility. A Don’t Know option was also available.

Ethical approval

This study was approved by the Medical Sciences Inter Divisional Research Ethics Committee of the University of Oxford (study ID:MSD-IDREC-C1-2014-152).

Statistical Analysis

SPSS Version 22 (Chicago, IL) was used to perform data analysis. Inter-rater agreement was measured by linear-weighted Cohen’s kappa between the two assessors. Non-parametric tests were used for the total GRS scores and metrics from the simulator reports as these data were not normally distributed. The number of performance metrics reported by the simulator which are able to distinguish between groups is reported as a percentage of the total number of metrics of that kind recorded. Kruskal-Wallis tests were performed for the total ASSET score and for each metric reported by the simulator, grouped by experience level. If statistically significant differences were found, pairwise Mann-Whitney U tests between the Novice-Intermediate, Intermediate-Expert and Novice-Expert groups to identify the levels at which the various metrics could distinguish between the groups. A p-value of <0.05 was taken to be significant in each case.

Other descriptive statistics for these data were reported as median values, with lower and upper quartiles. Face validity Likert scale responses were treated as ordinal, and reported as percentages of responses in agreement with statements[20].

Results

Inter-rater agreement

There was good agreement of the ASSET global rating scale results between the raters (KA and MB). Cohen’s linear-weighted kappa coefficient of agreement was 0.8519 (95% C.I.: 0.6918 – 1).

Construct validity

Across both the knee and shoulder groups, a total of six simulated tasks were assessed. The ASSET global rating scale showed statistically significant differences ($p < 0.00003$) between the three experience groups for each of the tasks. Subsequent pairwise analysis demonstrated statistically significant differences between both the

Novice and Intermediate group, and the Intermediate and Expert groups, for each task ($p \leq 0.003$ in all tasks). (Table 4 & Figure 2)

Of the 18 established performance metrics measured by the simulator, 14 (78%) could distinguish participants based on arthroscopic experience. The measured time taken to perform each task showed statistically significant differences between experience groups in all cases ($p \leq 0.001$). Subsequent pairwise analysis showed statistically significant differences between ten of the 12 pairs ($p \leq 0.01$): pairwise differences were not significant between Novice and Intermediate groups in the Knee Meniscectomy II task nor the Intermediate and Expert group in the Shoulder Guided Diagnostics I task. Path length was measured for 12 instruments across six tasks, and was able to distinguish participants in two-thirds of cases ($p \leq 0.007$).

Novel metrics include some measured metrics, and all of the scores calculated from measured values. Across the six tasks, only 5 of the 14 (36%) measured values for novel performance metrics could distinguish between participants based on arthroscopic experience. A total of 39 scores were calculated from the various established and novel measured values. Twenty-one of 39 scores (54%) showed significant differences between experience groups. 'Scratches' were measured as a percentage of proximal (femur/glenoid) and distal (tibia/humerus) joint surface area. Pairwise analysis demonstrated significant differences between intermediate and expert groups only ($p \leq 0.044$) in the knee tasks. No significant differences were found during the shoulder tasks between experience groups for glenoid or humerus scratch percentages. The Guided Meniscectomy II task reported an additional two novel measured values relating to 'Optimal medial region' (the percentage of meniscal pathology debrided) and 'Healthy medial region' (the percentage of normal meniscus inadvertently removed) (see Figure 1). Only the measured value for the 'Optimal medial region' was able to distinguish participants based on experience ($p \leq 0.009$).

Face validity

The questionnaire results are presented in Figure 3. Participants generally *agreed* (4) or *strongly agreed* (5) with statements regarding the realism of the external appearance, the display and the instrumentation were supported (93.6%, 87.1% and 93.6%, respectively). However, the realism of the feel of the bone and soft tissues (the key passive haptic features of this simulator), were not well supported (51.6% and 32.3%, respectively). The simulated diagnostic arthroscopy was realistic (90.3%), gave a sense of what it would be like to perform a real arthroscopy (91.7%), and provides both an enjoyable (97%) and unthreatening (90.9%) learning environment. It was felt the simulator would be useful for early years (100%) more than middle grade (78.9%) trainees, but only one experienced arthroscopic surgeon agreed it would be of benefit to consultants (16.7%).

Discussion

The most important finding of the present study was that this simulator could be both objectively useful at distinguishing candidates, and subjectively useful to trainees as a training tool. Of the 53 task modules available within this particular simulator, we assessed the tasks most relevant to everyday basic arthroscopy, testing the important skills of bimanual dexterity, orientation, and triangulation. There was strong construct validity when using a validated GRS to assess the variety of simulated tasks tested. This was also the case when using the established output metrics of the simulator, but a number of the novel metrics generated by the simulator were not useful in distinguishing between surgical experience levels. While the simulator also displays generally good face validity (Figure 3) and is seen as a useful training tool for early and middle years training, this version (1.2) did not rate well with regard to the realism of the passive haptic feedback which is one of its key features. This study extends previous work[13], being from an independent centre and assessing a broader range of tasks in both knee and shoulder simulation with a greater variety of metrics.

Rasmussen describes three levels of human behaviour: skill-, rule- and knowledge-based[30], and the selection of guided VR tasks (where the simulator advises of the next steps in the task) was deliberate to allow this study to focus on the skill-based behaviour measured by the simulator and not be confounded by the cognitive aspects of the procedure which may be unfamiliar, especially to novices[38]. In real world surgical performance, the rule- and knowledge-based behaviours in decision making during surgery are significant[10, 23], and the tasks selected may dampen that effect. Pragmatically, only some of the available 53 tasks could be assessed, and while the six selected by shoulder and knee arthroscopy experts were best matched to the UK orthopaedic curriculum[6], if more could have practically have been assessed it would have given a more complete insight into the ArthroS simulator.

The presence of haptic feedback can improve performance of training on simulators and is felt to justify the additional costs[28]. This study agrees with previous findings[13] that the tactile feedback was not felt to be that realistic, with mid-scale Likert scores for face validity regarding the bone and soft tissue tactile feedback (Figure 3). It important to note that face validity is only a superficial impression of the device and often felt to be of lesser importance in the overall evaluation of simulators[12, 37]. The most realistic passive haptic simulation of surgical tissues is from fresh frozen cadavers or live anaesthetized animals, but both are costly and have ethical implications[7]. Though many efforts to improve simulation focus on the ‘realism’ with introduction of features such as advancing VR techniques, passive haptics, and realistic joints with convincing articulation, these improvements are associated with increasing costs which is not met by more funding for training programmes[26]. In response to this, low fidelity arthroscopic simulator environments have been explored which show construct validity[29]. It must be recognized that the cost of increased fidelity, including haptics, may be unaffordable given that low fidelity simulators seem as effective[9, 22], and efforts to further improve the passive haptic feedback should be redirected at other aspects of simulator design.

By using a GRS, the present study provides a more detailed assessment of construct validity by using an established gold standard measure for concurrence alongside the simulator reported metrics. Interestingly, while some new novel metrics were proposed in the previous study[13] and used by this version of the simulator, they currently fail to reliably differentiate between experience levels for 9 of the 14 (64%) metrics assessed. While the current study used the Version 1.2 software, it has since been further updated to try and address these some of these limitations.

The simulator provides “expert-defined scoring”[36] (see Figure 1), but the usefulness of these scores to the user remains unclear. Only 54% of these could distinguish participants based on experience in in this study. The scores are calculated from the measured metrics e.g. time taken, path length, but add little extra information to the values from which they are calculated. In 95% of cases, the calculated metric simply confirmed the difference between groups found in the measured metric.

The established and recognized performance metrics were twice as likely (78% vs 36%) to show a statistically significant difference between experience groups than the novel measured metrics. Despite this, it could be argued that the novel metrics help add more context and relevance to the user experience, even though at present, many cannot distinguish between levels of surgical expertise. For example, it does seem sensible to attempt to measure iatrogenic cartilage damage, but the manner in which this novel metric is derived at present needs further refinement to ensure usefulness. Other devices, such as PASSPORT v2, detect force at the level of the joint as a marker of safety[32, 34]: though ArthroS does not quantify the force during the simulation, it produces an image of the locations where excessive force has been applied when the task is complete to show the distribution of potential injury to the candidate. Future software updates addressing these issues including defining which metrics are ‘user experience metrics’ and which metrics (or combination of metrics) are ‘performance and learning metrics’ would be useful.

This study may be limited by group sizes and the choice of tasks in the study. There is disparity in the group sizes, however, this is not uncommon in construct validity studies[3, 13, 19, 21, 32, 34]. In fact, because of the differing spread of ability within experience groups, larger groups for novices and smaller for experts is desirable. There is known to be a greater variability in the innate ability of novices[3], so a larger sample size is better able represent this population[32]. Because proficiency measurement tools such as the ASSET instrument are designed to measure up to a level defined as expert[21], sampling a larger number of experts will be of limited benefit due to the built-in ceiling effect. This does not add bias: the simulator is not being used to define the skill of the participants, but participants with expected skill differences are being used as a standard to test if the simulator detects these skill differences. Only 6 of the available 53 tasks were selected for this study. This might lead to the criticism that if other tasks were used it may have affected the construct validity (particularly with regards to the novel metrics), or the face validity (particularly the realism of the passive haptics). A small sample of the available tasks was selected for two reasons. Firstly, it would be impractical to test all participants on all tasks, and so expert opinion was used to select those most representative of the skills needed for arthroscopy. Secondly, with a large number of tasks, participants attempts at earlier tasks during the study could be seen as ‘practice’ and might affect their performance on the following tasks. It risks mixing testing and training, and we might have seen a ‘within study training effect’ which would disproportionately affect the novice group. For that reason the number of attempts needed to be limited.

Developing objective metrics is the key to effective simulation training[5, 10]. Simulation researchers are looking at ‘novel metrics’ such as instrument loss and triangulation time which have been used as simple measures to assess performance and have been shown to correlate well with GRS[4]. Combining novel metrics

with the many established metrics which have been well validated in the orthopaedic and surgical literature[11, 16] would seem to be worthy goal, as they may well provide more richness to performance assessment but they need to be useful and validated parameters. A number GRS's have been developed[15], but they are time consuming to complete and require training and direct supervision to deliver. If a GRS could be generated automatically by a simulator, it would relieve the need for an assessor, reducing costs and potentially making simulation training more accessible.

Clinically, departmental funds invested in simulation and time spent training on a simulator aims to result in better performance by trainees in theatre. Evidence from other surgical sub-specialties suggests that VR based simulation training can translate to real differences in real theatre performance[27]. As such, it is prudent that orthopaedics follows suit, and develops effective evidence based simulators as this will have a positive impact on patient safety and care.

Conclusion

This study demonstrates the construct validity of an updated VR simulator with new passive haptic feedback using a gold standard instrument (the 'ASSET' global rating scale) and the simulator's output metrics over a variety of clinically relevant exercises. Despite poor realism of the passive haptic feedback, trainees feel this simulator is useful, and further development of the haptics may not improve its utility to trainees. Refining novel, construct valid, clinically appropriate metrics should be the focus of simulator development.

Acknowledgements

The ArthroS simulator used in this study was provided on loan from VirtaMed. The NIHR Oxford Musculoskeletal Biomedical Research Unit provided infrastructure support.

REFERENCES:

1. Aggarwal R, Tully A, Grantcharov T, Larsen CR, Miskry T, Farthing A, Darzi A (2006) Virtual reality simulation training can improve technical skills during laparoscopic salpingectomy for ectopic pregnancy. *Brit J Obstet Gynaec* 113:1382–1387
2. Akhtar K, Standfield NJ, Gupte CM, Tuijthof GJM (2015) Chapter 7: Virtual Reality Simulators. In: Karahan M, Kerkhoffs GMMJ, Randelli P, Tuijthof GJM (eds) *Effective Training in Arthroscopic Skills*, 1st edn. Springer Berlin Heidelberg, pp 71–80
3. Alvand A, Auplish S, Khan T, Gill HS, Rees JL (2011) Identifying orthopaedic surgeons of the future: the inability of some medical students to achieve competence in basic arthroscopic tasks despite training: a randomised study. *J Bone Joint Surg Br* 93:1586–1591
4. Alvand A, Khan T, Al-Ali S, Jackson WF, Price AJ, Rees JL (2012) Simple visual parameters for objective assessment of arthroscopic skill. *J Bone Joint Surg Am* 94:e97 doi: 10.2106/JBJS.K.01437
5. Angelo RL, Ryu RKN, Pedowitz RA, Gallagher AG (2015) Metric Development for an Arthroscopic Bankart Procedure: Assessment of Face and Content Validity. *Arthroscopy* 31:1430–1440
6. Frostick S, Baird E, Bale S, Banks T, Bhowal B, Kellett C, Cole A, Goodwin M, Hadfield-Law L, Hopgood P, Pitts D, Turner P, Reed M, Sher L, Tudor F (2013) *Trauma and Orthopaedic Curriculum Mapped to Simulation Options*. British Orthopaedic Association, London. Available at https://www.iscp.ac.uk/static/public/TO_Putting_simulation_into_practice.pdf
7. Brown D (2013) The Role of Simulation in the Learning of Surgical Skills. *Ann R Coll Surg Eng (Suppl)* 95:292–295
8. Cannon WD, Nicandri GT, Reinig K, Mevis H, Wittstein J (2014) Evaluation of skill level between trainees and community orthopaedic surgeons using a virtual reality arthroscopic knee simulator. *J Bone Joint Surg Am* 96:e57 doi: 10.2106/JBJS.M.00779
9. Coughlin RP, Pauyo T, Sutton JC, Coughlin LP, Bergeron SG (2015) A Validated Orthopaedic Surgical Simulation Model for Training and Evaluation of Basic Arthroscopic Skills. *J Bone Joint Surg Am* 97:1465–1471
10. Darzi A, Smith S, Taffinder N (1999) Assessing operative skill: needs to become more objective. *Brit Med J* 318:887–888
11. Datta V, Mackay S, Mandalia M, Darzi A (2001) The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. *J Am Coll Surg* 193:479–485
12. Ferguson JY, Alvand A, Price AJ, Rees JL (2015) Chapter 8: Theory on Simulator Validation. In: Karahan M, Kerkhoffs GMMJ, Randelli P, Tuijthof GJM (eds) *Effective Training in Arthroscopic Skills*, 1st edn. Springer Berlin Heidelberg, pp 81–94
13. Fucentese SF, Rahm S, Wieser K, Spillmann J, Harders M, Koch PP (2015) Evaluation of a virtual-reality-based simulator using passive haptic feedback for knee arthroscopy. *Knee Surg Sports Traumatol Arthrosc* 23:1077–1085
14. Gélinas-Phaneuf N, Choudhury N, Al-Habib AR, Cabral A, Nadeau E, Mora V, Pazos V, Debergue P, DiRaddo R, Del Maestro RF (2014) Assessing performance in brain tumor resection using a novel virtual reality simulator. *Int J Comput Assist Radiol Surg* Springer Berlin Heidelberg 9:1–9 doi: 10.1007/s11548-013-0905-8
15. Hodgins JL, Veillette C (2013) Arthroscopic proficiency: methods in evaluating

- competency. *BMC Med Educ* 13:61
16. Howells NR, Brinsden MD, Gill RS, Carr AJ, Rees JL (2008) Motion Analysis: A Validated Method for Showing Skill Levels in Arthroscopy. *Arthroscopy* 24:335–342
 17. Howells NR, Gill HS, Carr AJ, Price AJ, Rees JL (2008) Transferring simulated arthroscopic skills to the operating theatre: a randomised blinded study. *J Bone Joint Surg Br* 90:494–499
 18. Hunter S, McLaren P (1993) Specialist medical training and the Calman report. *Brit Med J* 306:1281–1282
 19. Jacobsen ME, Andersen MJ, Hansen CO, Konge L (2015) Testing Basic Competency in Knee Arthroscopy Using a Virtual Reality Simulator: Exploring Validity and Reliability. *J Bone Joint Surg Am* 97:775–781
 20. Jamieson S (2004) Likert scales: how to (ab)use them. *Med Educ* 38:1217–1218
 21. Koehler RJ, Amsdell S, Arendt EA, Bisson LJ, Bramen JP, Butler A, Cosgarea AJ, Harner CD, Garrett WE, Olson T, Warne WJ, Nicandri GT (2013) The Arthroscopic Surgical Skill Evaluation Tool (ASSET). *Am J Sports Med* 41:1229–1237
 22. Lopez G, Wright R, Martin D, Jung J, Bracey D, Gupta R (2015) A Cost-Effective Junior Resident Training and Assessment Simulator for Orthopaedic Surgical Skills via Fundamentals of Orthopaedic Surgery: AAOS Exhibit Selection. *J Bone Joint Surg Am* 97:659–666
 23. Marsh H (2015) Better not look down.... *Ann R Coll Surg Eng (Suppl)* 97:339–342
 24. Milburn JA, Khera G, Hornby ST, Malone PSC, Fitzgerald JEF (2012) Introduction, availability and role of simulation in surgical education and training: Review of current evidence and recommendations from the Association of Surgeons in Training. *Int J Surg* 10:393–398
 25. Morrow G, Burford B, Carter M, Illing J (2012) The Impact of the Working Time Regulations on medical education and training: Final report on primary research. General Medical Council, London. Available at http://www.gmc-uk.org/The_Impact_of_the_Working_Time_Regulations_on_Medical_Education_and_Training_Final_Report_on_Primary_Research.pdf_51157039.pdf.
 26. Nousiainen MT, McQueen SA, Ferguson P, Alman B, Kraemer W, Safir O, Reznick R, Sonnadara R (2015) Simulation for Teaching Orthopaedic Residents in a Competency-based Curriculum: Do the Benefits Justify the Increased Costs? *Clin Orthop Relat Res* doi: 10.1007/s11999-015-4512-6
 27. Palter VN, Grantcharov TP (2014) Individualized Deliberate Practice on a Virtual Reality Simulator Improves Technical Performance of Surgical Novices in the Operating Room. *Ann Surg* 259:443–448
 28. Panait L, Akkary E, Bell RL, Roberts KE, Dudrick SJ, Duffy AJ (2009) The Role of Haptic Feedback in Laparoscopic Simulation Training. *J Surg Res* 156:312–316
 29. Pedowitz RA, Nicandri GT, Angelo RL, Ryu RKN, Gallagher AG (2015) Objective Assessment of Knot-Tying Proficiency With the Fundamentals of Arthroscopic Surgery Training Program Workstation and Knot Tester. *Arthroscopy* 31(10):1872–1879
 30. Rasmussen J (1983) Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE T Sys Man Cyb* 13:257–266
 31. Shah J, Darzi A (2002) Virtual reality flexible cystoscopy: a validation study. *Brit J Urol.* 90:828–832
 32. Stunt JJ, Kerkhoffs GMMJ, Horeman T, van Dijk CN, Tuijthof GJM (2014) Validation of the PASSPORT V2 training environment for arthroscopic skills. *Knee Surg Sports Traumatol Arthrosc* doi: 10.1007/s00167-014-3213-0
 33. Stunt JJ, Kerkhoffs GMMJ, van Dijk CN, Tuijthof GJM (2015) Validation of the

- ArthroS virtual reality simulator for arthroscopic skills. *Knee Surg Sports Traumatol Arthrosc* 23:3436–3442
34. Tashiro Y, Miura H, Nakanishi Y, Okazaki K, Iwamoto Y (2009) Evaluation of skills in arthroscopic training based on trajectory and force data. *Clin Orthop Relat Res* 467:546–552
 35. Tuijthof GJM, van Sterkenburg MN, Sierevelt IN, van Oldenrijk J, Van Dijk CN, Kerkhoffs GMMJ (2010) First validation of the PASSPORT training environment for arthroscopic skills. *Knee Surg Sports Traumatol Arthrosc* 18:218–224
 36. VirtaMed AG (2014) VirtaMed ArthroS Virtual reality training simulator for knee and shoulder arthroscopy factsheet. VirtaMed, Zurich. Available at http://www.virtamed.com/files/1914/2504/7801/VirtaMed_ArthroS_Factsheet_150227.pdf
 37. Weiner IB, Craighead WE (Eds.) (2010) *The Corsini Encyclopedia of Psychology*. Wiley, Hoboken, New Jersey
 38. Wentink M, Stassen LPS, Alwayn I, Hosman RJAW, Stassen HG (2003) Rasmussen's model of human behavior in laparoscopy training. *Surg Endosc* 17:1241–1246
 39. Ziegler R, Fischer G, Müller W, Göbel M (1995) Virtual Reality Arthroscopy Training Simulator. *Comput Biol Med* 25:193–203

Figure 1 - tiff format

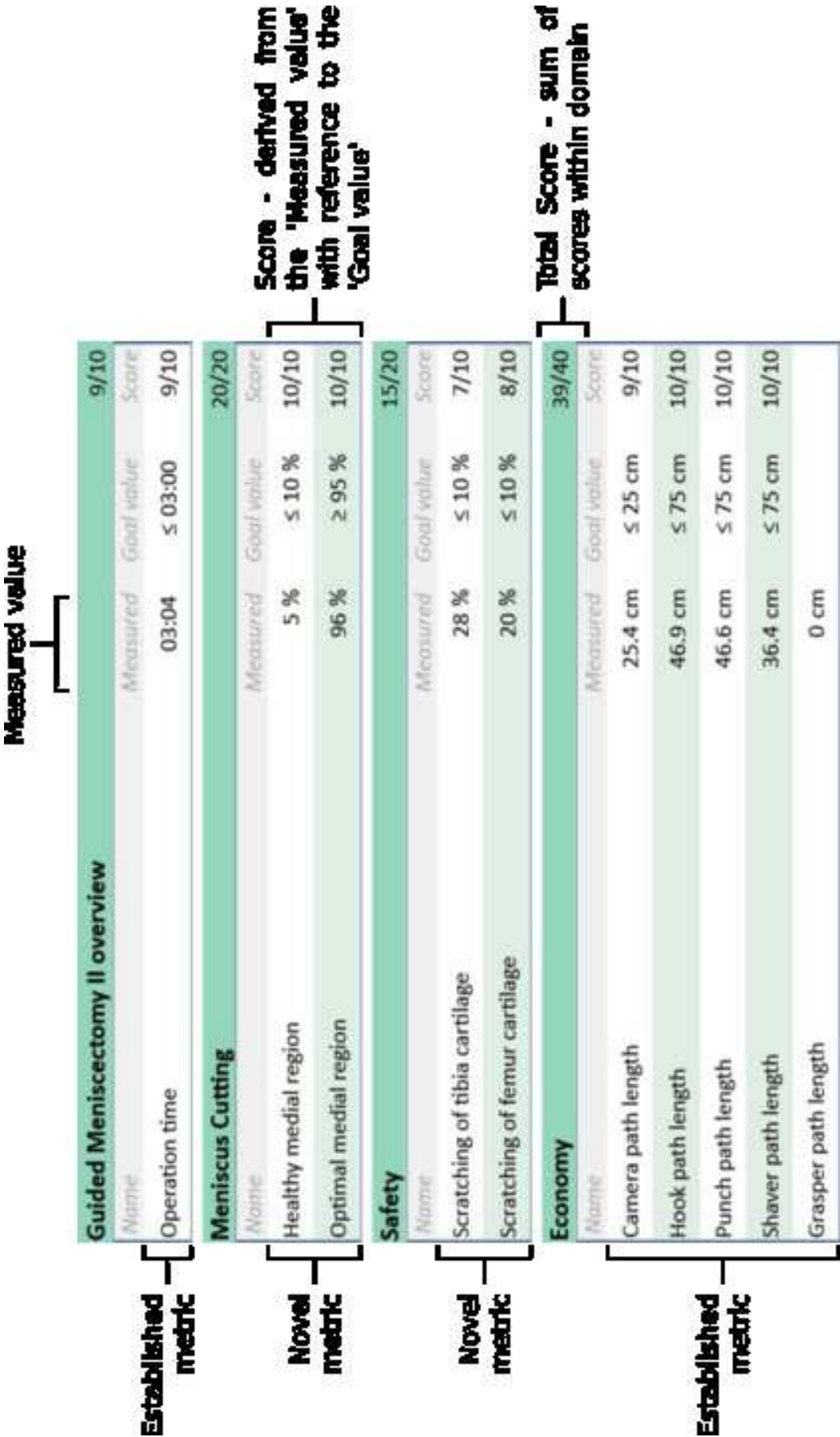


Figure 2a - eps format

Fig 2a: Performance variation between groups during Knee Guided Diagnostics II task

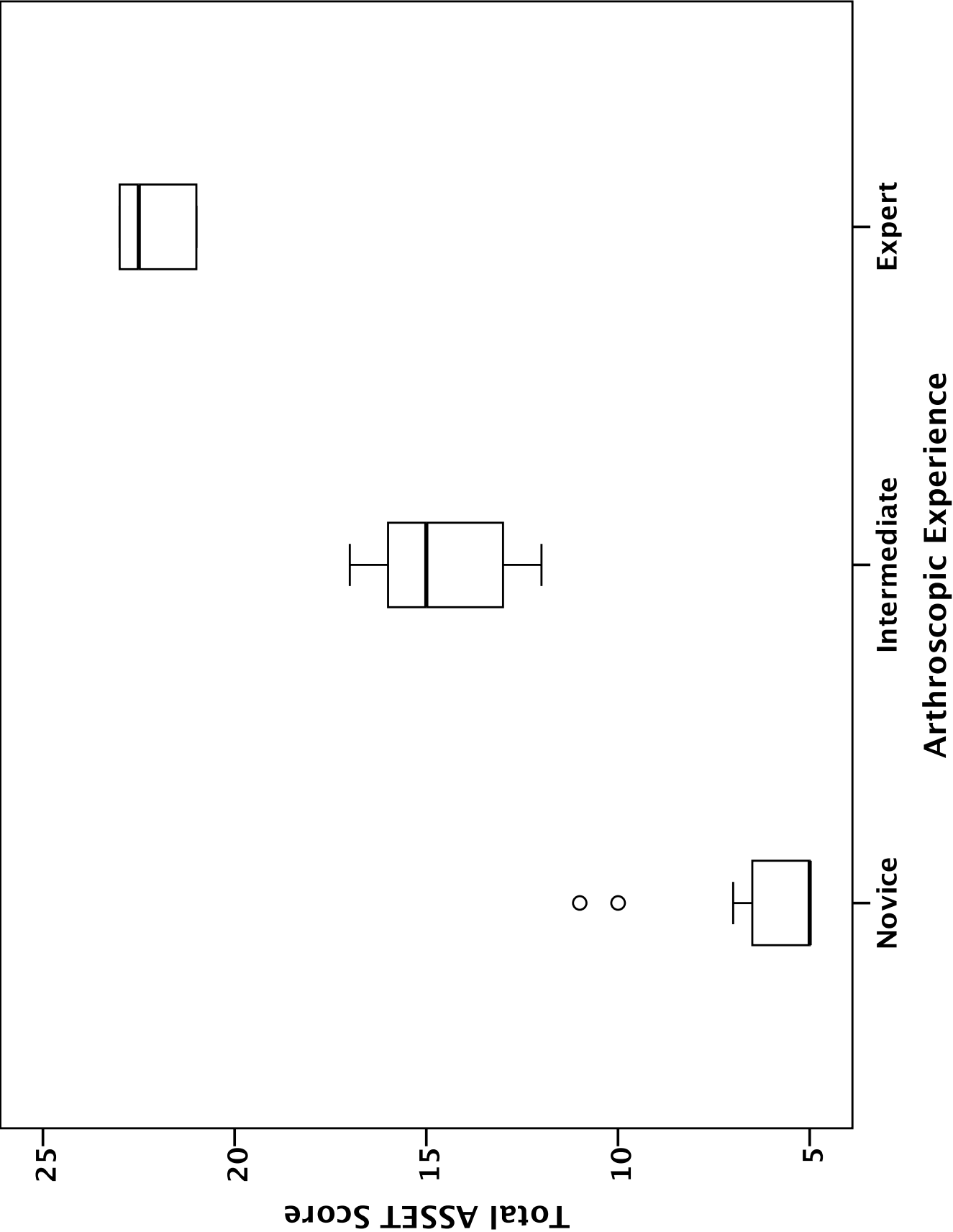


Figure 2b - eps format

Fig 2b: Performance variation between groups during Shoulder Guided Diagnostics I task

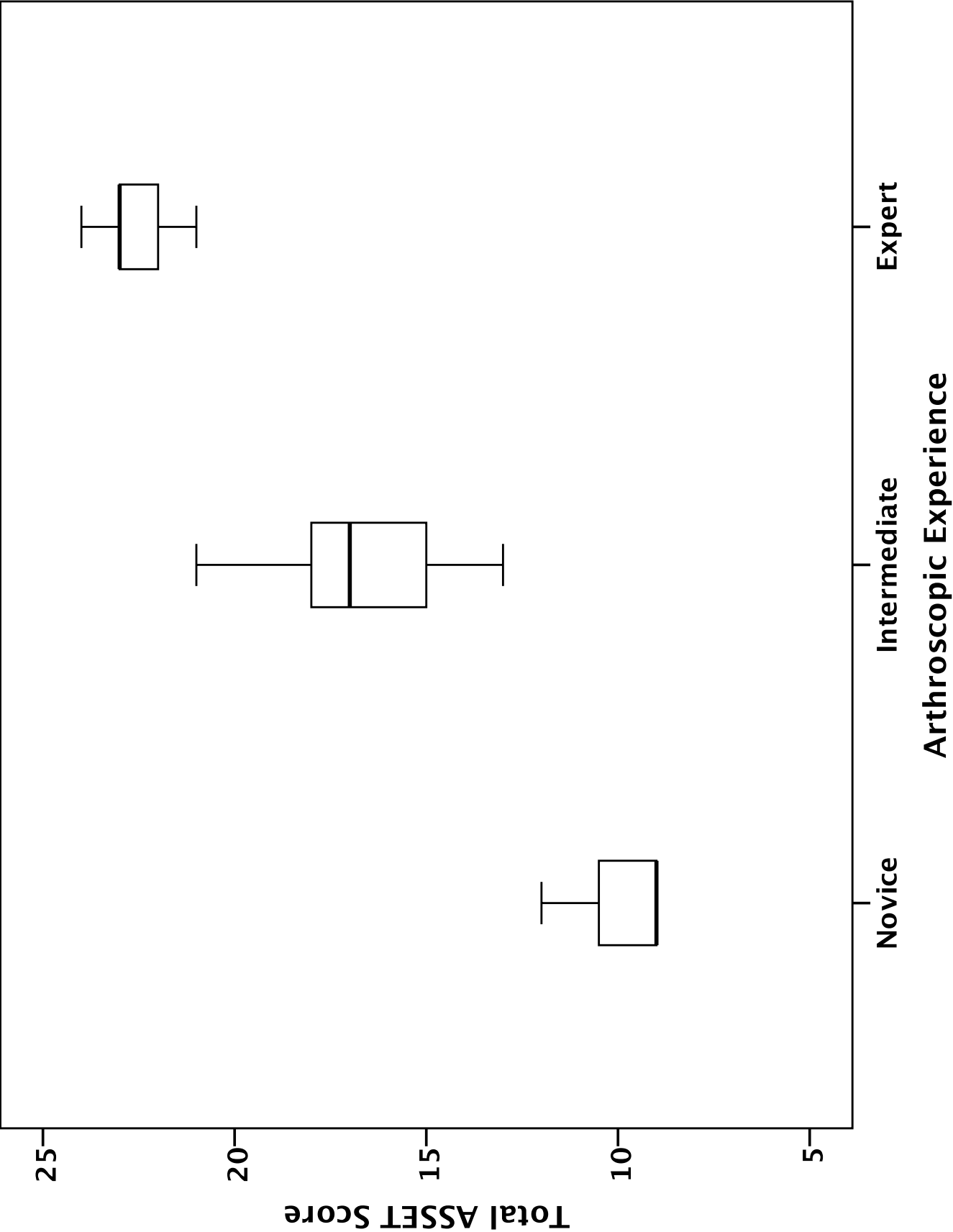


Figure 3 - eps format

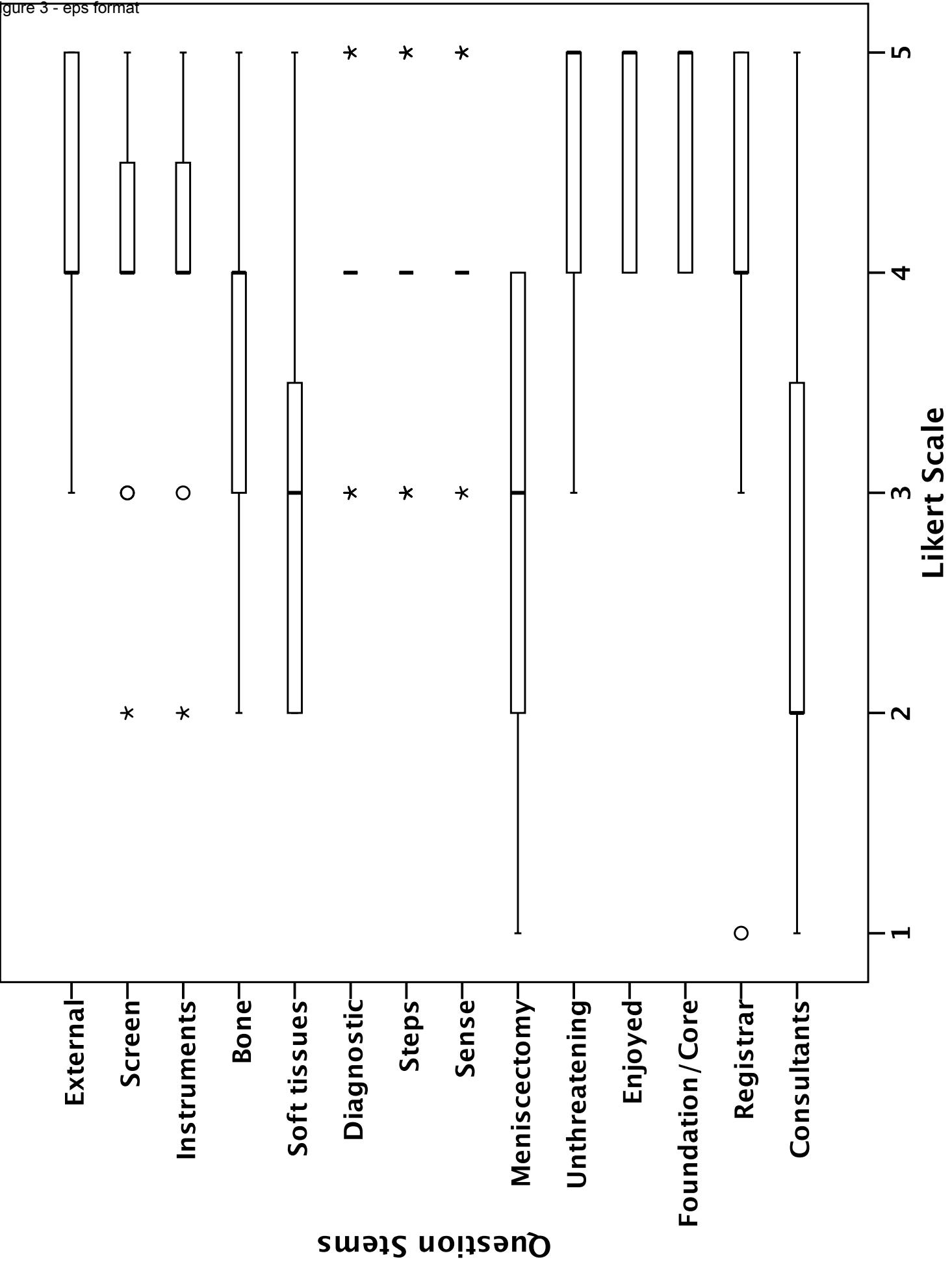


Figure 1 - Annotated example of a simulator report: The metrics produced by the report have been labelled to indicate if they are established (i.e. have been validated in other studies), or novel to this simulator. Each metric is reported as a *Measured* value, and as a *Score* out of 10. The *Goal value* is required to score maximum points. All scores are novel metrics. Descriptions of metrics: *Operation time* - time in minutes and seconds to complete the task. *Healthy medial region* - Percentage of normal meniscus removed. *Optimal medial region* - Percentage of meniscal lesion removed. *Scratching* - percentage of cartilage surface area damaged by arthroscopic instruments during task. *Path length* - distance described by the instruments during the task.

Figure 2 - Box-whisker plots showing construct validity of knee (a) and shoulder (b) diagnostic tasks using ASSET

Figure 3 - Summary of the face validity questionnaire responses. Likert scale anchored to the statements 1:Strongly disagree, 2:Disagree, 3:Neither agree or disagree, 4:Agree, 5:Strongly agree. Question stems: *External* – The external instrumentation was realistic, *Screen* – The visual experience of the arthroscopy screen was realistic, *Instruments* – the visual appearance of the instruments on screen was realistic, *Bone* – the feel of the bone was realistic, *Soft tissues* – the feel of the soft tissues was realistic, *Diagnostic* – the arthroscopy procedures was realistic, *Steps* - the steps performed in the simulator accurately reflect the steps taken in the actual procedure, *Sense* – the simulator gave a sense of what a real arthroscopy would be like, *Meniscectomy* – the arthroscopic meniscectomy was realistic, *Unthreatening* – the simulator provided and unthreatening learning environment, *Enjoyed*, - I enjoyed using the simulator, *Foundation/Core* – the simulator is a useful training tool for foundation and core trainees (equivalent to intern level), *Registrar*, - the simulator is a useful training tool for registrars (equivalent to resident level), *Cons* – the simulator is a useful training tool for consultants (equivalent to attending level).

Table 2

	Knee Simulation			Shoulder Simulation		
	Novice	Intermediate	Expert	Novice	Intermediate	Expert
Number of participants	15	10	5	15	10	5
Number of Arthroscopies performed	0	42.4 (12-90)	760.5 (400-1000)	0	23.1 (1-50)	1288.2 (150-3000)

Table 2 – Table showing arthroscopic experience of the participants in knee and shoulder simulation groups (mean number of arthroscopies performed; range in brackets)

Table 3

	Knee group (30 participants: 15 novice, 10 intermediate, 5 expert)	Shoulder group (30 participants: 15 novice, 10 intermediate, 5 expert)
Task One	<i>Guided Diagnostic II</i> - visual examination of 13 landmarks (Patella, Suprapatellar pouch, Popliteus, Trochlea, Lateral meniscus posterior horn, Lateral meniscus intermedia, Lateral meniscus anterior horn, Medial Meniscus posterior horn, Medial meniscus intermedia, Medial meniscus anterior horn, Posterior cruciate, Distal anterior cruciate, Proximal anterior cruciate), and probing of posterior horn medial meniscal tear	<i>Guided Diagnostics I</i> - visual examination of 10 landmarks (Biceps tendon, Supraspinatus, Infraspinatus, Subscapularis, Humerus, Glenoid cartilage, Dorsal labrum, Superior labrum, Anterior medial labrum, Inferior labrum)
Task Two	<i>Triangulation III</i> - hook 6 rings with probe	<i>Triangulation I</i> - probe examination of 5 spheres
Task Three	<i>Guided Meniscectomy II</i> - Visualise a posterior horn medial meniscal tear, perform a meniscectomy and assess stability of the operated meniscus	<i>Triangulation III</i> - hook 5 rings with probe

Table 3 – Table showing the checklists of steps to be performed during simulated knee and shoulder arthroscopic tasks

Validation	Description
Content	<i>The simulation contains the important components of a task as assessed by an expert; this represents the educational content of a simulator</i>
Face	<i>The impression of non-expert end-users about how realistic a simulation feels ('haptics') and looks; this represents the acceptability of a simulator to trainees</i>
Construct	<i>The simulator can accurately differentiate candidates based on relative surgical skill; this represents the ability of a simulator to be used as an assessment tool</i>
Transfer	<i>Training on a simulator improves performance in the operating theatre; this represents the effect of simulation on real world performance</i>

Table 1 - Forms of simulator validity

Table 4 revised

		<i>p-value, Across all three groups[†]</i>	Novice group	<i>p-value, Novice vs Intermediate^{††}</i>	Intermediate group	<i>p-value, Intermediate vs Expert^{††}</i>	Expert group	<i>p-value, Novice vs Expert^{††}</i>
Knee	Task One – Guided Diagnostics II	0.000001	5 (5-6.5)	0.00001	15 (13-16)	0.001	22.5 (21-23)	0.0002
	Task Two - Triangulation III	0.000004	8 (7.5-10)	0.00006	20 (18-21)	0.001	29 (28-32)	0.0005
	Task Three - Guided Meniscectomy II	0.000003	12 (10.5-15.5)	0.00004	23 (22-25)	0.001	31 (31-31)	0.0005
Shoulder	Task One – Guided Diagnostics I	0.000004	9 (9-10.5)	0.00003	17(15-18)	0.003	23 (22-23)	0.0008
	Task Two - Triangulation I	0.000005	13 (11-16)	0.00003	22 (22-23)	0.002	33 (32-35)	0.001
	Task Three - Triangulation III	0.00003	12 (8.5-13)	0.0004	19.5 (18-22)	0.003	33 (28-34)	0.001

Table 4 – Table showing differences in mean total ASSET scores for each task, grouped by experience level. Differences between groups were all statistically significant on pairwise analysis. [†]Kruskal-Wallis Test ^{††}Mann-Whitney-U Test.