

Improving simulation training in orthopaedics

Patrick Garfjeld Roberts

Hertford College
University of Oxford

*A thesis submitted for the degree of
Doctor of Philosophy*

Hilary 2018

Abstract

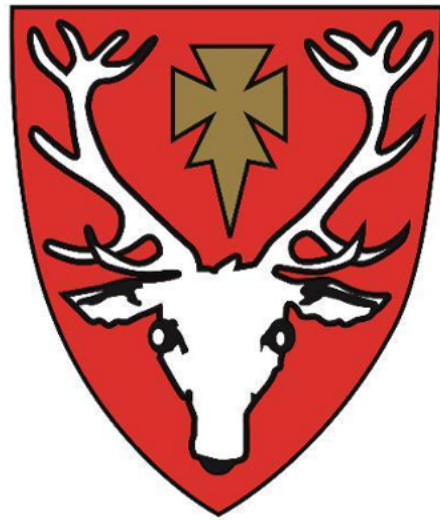
The way surgical trainees acquire technical skills is changing in modern surgical training programmes: simulation is proposed as a key part of those changes. Arthroscopy is a surgical technique that is increasing in both incidence and technical complexity; where simulation is becoming common, but evidence is limited. Real-world performance improvements can be measured following simulation training in other fields, but equivalent measures of intra-operative performance are inadequate. Thus, although surgical simulation is popular and improves simulated performance, there is little objective evidence that it improves intra-operative performance.

The original contribution of this thesis is to objectively demonstrate the transfer of simulation training into improved intra-operative technical skills. To achieve this, a systematic literature review investigated the quantitative metrics currently used to measure arthroscopic performance, identifying wireless motion analysis as a potential method to assess performance intra-operatively. Motion analysis is a recognised objective method to measure surgical activity which correlates with surgical experience, so wireless motion analysis was validated against a wired motion analysis method commonly used in simulation but not feasible for intra-operative use. Wireless motion analysis metrics were further validated with a simulated arthroscopy list: this environment allowed deliberate practice of arthroscopic sub-skills with proximate feedback for independent practice.

This simulated arthroscopy list with wireless motion analysis was used in two randomised studies: the penultimate study of this thesis investigated the impact of simulated practice on the arthroscopic learning curve and showed that performance improved rapidly with independent practice but was not modified by feedback, while the final study investigated additional simulation practice during early surgical training, and objectively demonstrated that additional simulation training improved intra-operative performance compared to traditional training alone.

This thesis is the first to objectively show that simulation affects intra-operative behaviour. It sets the groundwork for further investigations into efficient, cost-effective simulation and the impact of simulation training on patient outcomes.

Improving simulation training in orthopaedics



Patrick Garfjeld Roberts

Hertford College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary 2018

Acknowledgements

I would like to thank my supervisors, Professor Jonathan Rees and Mr Abtin Alvand. A decade ago, as a medical student, I was a participant in an early simulation study at OOSEC which was published under the title “Identifying orthopaedic surgeons of the future: the inability of some medical students to achieve competence in basic arthroscopic tasks despite training”. I had developed an interest in medical education in clinical school and continued this into my surgical career, when some years later Abtin called me to say the boss had something to discuss with me: this project stemmed from that call. I thank them both for thinking of me, giving me this opportunity, and I hope to continue working with them in the future. But I have never discovered if I was one of those untrainable students...

This project rests on the impact simulation has in the operating theatre. I would like to thank the consultants from the Acute Knee Service at the Nuffield Orthopaedic Centre: Professor Andrew Price, Mr Nick Bottomley, Mr Mark Deakin, and Mr Will Jackson. They allowed me into their theatres, and I am grateful for their commitment to the project. I thank the participants in the studies too: I spent many hours in the skill lab with some of them, and I wish them well with their future careers, surgical or otherwise.

The Botnar Research Centre is a world-leading research institute, and I am very grateful to Professor Andrew Carr for hosting me and the funding received via the NIHR Biomedical Research Centre. I had the pleasure of spending time and sharing ideas with many people at the Botnar, especially the orthopaedic research fellows from both Gryffindor and Slytherin who became great friends too. I would like to thank Professor Chris Lavy and Grace Le from the Global Surgery group, for their company, coffee, and letting me occupy their territory.

The Dinwoodie Charitable Company and the Royal College of Surgeons of England kindly funded the first two years of this project. Professor John Black and the Dinwoodie trustees have embraced medical education and simulation, and were always enthused and perceptive when we met to discuss the work. The College, and Martyn Coomer in particular, are committed to education and training, and I thank them for their faith in this project.

This project was a part of a collaboration with McLaren Applied Technologies, who were a pleasure to work with and enlightened me to the importance of making

academic research have a real-world impact. I would like to thank Professor Caroline Hargrove, Sami Khan, Dr Marco Gallieri, and Sinead Cook who have been an integral part of this collaboration, as well as all the others at MAT who were involved in the background.

Most importantly, I am thankful to my wife and two girls, Rachel, Cecilie, and Millie, and my parents and sister, Else and Brian, and Karin. They have always been there, and the past three years have been such a wonderful time for us as a family. Without them, this whole endeavour would have been neither possible nor worthwhile.

Abstract

The way surgical trainees acquire technical skills is changing in modern surgical training programmes: simulation is proposed as a key part of those changes. Arthroscopy is a surgical technique that is increasing in both incidence and technical complexity; where simulation is becoming common, but evidence is limited. Real-world performance improvements can be measured following simulation training in other fields, but equivalent measures of intra-operative performance are inadequate. Thus, although surgical simulation is popular and improves simulated performance, there is little objective evidence that it improves intra-operative performance.

The original contribution of this thesis is to objectively demonstrate the transfer of simulation training into improved intra-operative technical skills. To achieve this, a systematic literature review investigated the quantitative metrics currently used to measure arthroscopic performance, identifying wireless motion analysis as a potential method to assess performance intra-operatively. Motion analysis is a recognised objective method to measure surgical activity which correlates with surgical experience, so wireless motion analysis was validated against a wired motion analysis method commonly used in simulation but not feasible for intra-operative use. Wireless motion analysis metrics were further validated with a simulated arthroscopy list: this environment allowed deliberate practice of arthroscopic sub-skills with proximate feedback for independent practice.

This simulated arthroscopy list with wireless motion analysis was used in two randomised studies: the penultimate study of this thesis investigated the impact of simulated practice on the arthroscopic learning curve and showed that performance improved rapidly with independent practice but was not modified by feedback, while the final study investigated additional simulation practice during early surgical training, and objectively demonstrated that additional simulation training improved intra-operative performance compared to traditional training alone.

This thesis is the first to objectively show that simulation affects intra-operative behaviour. It sets the groundwork for further investigations into efficient, cost-effective simulation and the impact of simulation training on patient outcomes.

Contents

List of Figures	x
List of Tables	xii
1 Introduction	1
1.1 The practice of surgery	2
1.1.1 Ancient and pre-modern surgery	2
1.1.2 Domains of surgical skill	3
1.1.3 The role of technical skill	3
1.2 Arthroscopy	6
1.2.1 The evolution of arthroscopy	6
1.2.2 The psycho-motor challenges of arthroscopy	8
1.3 Training technical skills	9
1.3.1 Teaching technical skills	9
1.3.2 Learning technical skills	11
1.3.3 Assessing technical skills	14
1.4 Simulation	17
1.4.1 The roles of simulation in surgery	17
1.4.2 The case for simulation as a training intervention	18
1.4.3 Is surgical simulation effective?	20
1.5 Specific aims and objectives	24
1.5.1 Aim	24
1.5.2 Objectives	24
2 Systematic review	26
2.1 Introduction	27
2.2 Aims	28
2.3 Methods	28
2.3.1 Study identification	28
2.3.2 Registration	31
2.4 Results and analysis	32
2.4.1 Description of included studies	32

2.4.2	Quantitative methods of assessing arthroscopic performance	35
2.4.3	Time	35
2.4.4	Motion parameters	42
2.4.5	Global rating scales	46
2.4.6	Task-specific checklists	50
2.4.7	Error lists	53
2.4.8	Behavioural features	55
2.4.9	Force parameters	57
2.4.10	Single-item scales	59
2.4.11	Virtual reality proprietary metrics	61
2.4.12	Calculated scores	63
2.4.13	Fluoroscopy	66
2.4.14	Correlation of quantitative measures of arthroscopic performance to intra-operative behaviour	67
2.4.15	Correlation of quantitative measures of arthroscopic performance to patient-reported outcome measures	69
2.5	Discussion	69
2.5.1	Summary of findings	69
2.5.2	Strength and limitations	69
2.5.3	In the context of the wider literature	70
2.5.4	Implications of this study	71
3	Validating objective quantitative assessment of technical skill using WASP	74
3.1	Introduction	75
3.1.1	The role of motion analysis in the assessment of surgical skill	75
3.1.2	Validating new methods of assessment	76
3.1.3	Aims	78
3.2	Study A - concurrent validity of wired and wireless motion analysis	78
3.2.1	Methods	78
3.2.2	Results	84
3.3	Study B - accuracy and reliability of wired and wireless motion analysis	87
3.3.1	Methods	87
3.3.2	Results	89
3.4	Discussion	90
3.4.1	Implications of the studies	90
3.4.2	Limitations	95
3.4.3	Summary	96

4	Development of a simulated arthroscopy list	97
4.1	Introduction	97
4.1.1	Aim	99
4.2	Study A - Task selection for a simulated arthroscopy list	100
4.2.1	Methods	100
4.2.2	Results	105
4.3	Study B - Construct and face validity of a simulated arthroscopy list	108
4.3.1	Methods	108
4.3.2	Results	111
4.4	Discussion	122
4.4.1	Implications of the studies	122
4.4.2	Limitations	125
4.4.3	Summary	126
5	Randomised controlled trial to assess the impact of objective performance feedback on the learning curve during simulated arthroscopy practice	128
5.1	Introduction	129
5.1.1	Aim and hypothesis	131
5.2	Methods	132
5.2.1	Ethical approval	132
5.2.2	Participants and study setting	132
5.2.3	Simulation sessions, task, and feedback	132
5.2.4	Power calculation	134
5.2.5	LC-CUSUM	135
5.2.6	Outcomes and statistical analysis	136
5.3	Results	137
5.3.1	Baseline performance	137
5.3.2	Rate of learning: differences between groups each week . . .	139
5.3.3	Rate of learning: first attempt reaching ‘consultant level’ performance	139
5.3.4	LC-CUSUM	140
5.3.5	Performance change without knowledge of results	140
5.4	Discussion	143
5.4.1	Implications of the study	143
5.4.2	Limitations	144
5.4.3	Summary	147

6	Randomised controlled trial to objectively assess the transfer of surgical simulation training to intra-operative performance	148
6.1	Introduction	149
6.1.1	Aim and hypothesis	150
6.2	Methods	150
6.2.1	Ethical approval and registration	150
6.2.2	Participants and study setting	151
6.2.3	Equipment	151
6.2.4	Randomisation	152
6.2.5	Assessments and training	152
6.2.6	Outcome measures and statistical analysis	153
6.3	Results	155
6.3.1	Participant baseline simulated arthroscopy list performance .	155
6.3.2	Participant in-theatre performance	158
6.3.3	Supervisor in-theatre performance	158
6.3.4	Participant:supervisor in-theatre performance ratio	159
6.3.5	Participant follow-up simulated arthroscopy list performance	163
6.3.6	Participant training activities during study	163
6.4	Discussion	163
6.4.1	Implications of the study	163
6.4.2	Limitations	167
6.4.3	Summary	169
7	Discussion	171
7.1	Summary of main findings and conclusions	171
7.1.1	Contributions to the field of study	174
7.2	Clinical relevance	175
7.3	Limitations	176
7.4	Recommendations for future work	180
Appendices		
A	General appendices to methods for simulation studies	186
A.1	Arrangement of the skills laboratory	186
A.2	Equipment and instruments	187
A.3	Ethical approval	188
A.4	WASP motion analysis system	188
B	Systematic review of quantitative methods of assessing arthroscopic performance	195
B.1	Search strategy	196
B.2	Full tables of study characteristics	197

C	Development of a simulated arthroscopy list	205
C.1	FAST task selection survey	206
C.2	Meniscectomy task development	206
C.2.1	Process and materials	206
C.2.2	Meniscal tear morphology	208
C.3	Face and content validity survey	210
C.4	Construct validation results tables	212
D	Randomised controlled trial to assess the impact of objective motion performance feedback on the learning curve in simulated arthroscopy	221
E	Randomised controlled trial to objectively assess the transfer of surgical simulation training to intra-operative surgical skills performance	223
E.1	Approved in-theatre protocol	223
E.2	Simulated arthroscopy list and in-theatre performance results	228
	References	234

List of Figures

1.1	Modern arthroscopic equipment	7
1.2	The effect of radial distortion	9
1.3	RCS(Eng) Training Record	16
2.1	PRISMA flow diagram	33
3.1	Wireless elbow-worn motion sensors	76
3.2	The relationships between recorded motion data	77
3.3	Accuracy and reliability	77
3.4	External view of arthroscopic shoulder simulator	80
3.5	Shoulder simulator landmarks	81
3.6	Comparison of wireless and wired sensors	82
3.7	Correlation between wired and wireless (ACC) motion analysis systems.	85
3.8	Agreement between wired and wireless (ACC) system analysis systems.	85
3.9	Correlation between wired and wireless (ACC+GYRO) motion analysis systems.	86
3.10	Agreement between wired and wireless (ACC+GYRO) motion analysis systems.	86
3.11	3D-printed rig to generate stereotyped arthroscopic movements. . .	88
3.12	External view of the FAST workstation.	89
3.13	Reliability of the wired and wireless motion analysis systems during stereotyped movements.	90
3.14	Reliability variation with track length during stereotyped movements.	91
4.1	The relationships between the studies conducted to develop the simulated arthroscopy list	100
4.2	FAST arthroscopy workstation tasks	103
4.3	Excerpt from online FAST Task Selection survey	105
4.4	FAST task selection: difficulty.	106
4.5	FAST task selection: usefulness.	107
4.6	Anatomic tasks included in the simulated arthroscopy list.	110
4.7	Simulated arthroscopy list performance: time	113
4.8	Simulated arthroscopy list performance: hand movements	115

4.9	Simulated arthroscopy list performance: minor hand movements . . .	116
4.10	Simulated arthroscopy list performance: smoothness	118
5.1	WASP client ‘dashboard’	134
5.2	Learning curve: mean participant performance at each attempt . . .	138
5.3	LC-CUSUM plots	141
6.1	Agreed diagnostic knee arthroscopy sequence	154
6.2	CONSORT flow diagram	157
6.3	Simulated diagnostic knee arthroscopy performance: baseline & follow-up	158
6.4	Participant in-theatre diagnostic knee arthroscopy performance. . .	160
6.5	Supervisor in-theatre diagnostic knee arthroscopy performance. . .	161
6.6	Ratio of participant:supervisor in-theatre diagnostic knee arthroscopy performance.	162
7.1	SOQQ plot: everyday skill	184
7.2	SOQQ plot: surgical technical skill	184
A.1	Skills lab set-up for generic and anatomic bench-top simulation . . .	187
A.2	Arthroscopy equipment	189
A.3	Arthroscopy instruments	190
A.4	Ethical approval letter (MSD-IDREC-C1-2014-152)	191
A.5	Specifications of the WAX9 motion sensors	192
A.6	WAX9 sensor positioning	193
A.7	WASP client ‘dashboard’	194
B.1	Search strategy	196
C.1	Guides for creating reproducible bucket-handle and vertical oblique flap tears	207
C.2	Rejected sources of meniscal inserts	208
C.3	Taking a relief mould of a commercial meniscal insert.	209
C.4	Common meniscal tear morphologies.	210

List of Tables

1.1	Domains of surgical skill	4
1.2	Comparison of the Kirkpatrick levels with clinical trial phases. . . .	23
2.1	Keywords for title and abstract screening	29
2.2	Inclusion and exclusion criteria.	30
2.3	Inter-rater agreement between Reviewers A & B.	32
2.4	Characteristics of included studies.	36
2.5	Quantitative metrics in included studies	37
2.6	Evidence demonstrated for application of time as a quantitative performance metric.	37
2.7	Evidence demonstrated for application of motion parameters as quantitative performance metrics.	43
2.8	Subjectivity in global rating scales	48
2.9	Evidence demonstrated for application of global rating scales as a quantitative performance metric.	48
2.10	Published task-specific checklists	51
2.11	Evidence demonstrated for application of task-specific checklists as quantitative performance metrics.	52
2.12	Evidence demonstrated for application of error lists as quantitative performance metrics.	54
2.13	Evidence demonstrated for use of behavioural features as quantitative performance metrics.	56
2.14	Evidence demonstrated for application of force parameters as quanti- tative performance metrics	58
2.15	Evidence demonstrated for application of single-item scales as quan- titative performance metrics.	61
2.16	Evidence demonstrated for application of VR proprietary metrics as quantitative performance metrics.	62
2.17	Evidence demonstrated for application of calculated scores as quanti- tative performance metrics.	64
2.18	Risk of bias assessment	68
3.1	Study A participant characteristics	79

3.2	Location and sequence of numbered landmarks to probe during simulated diagnostic shoulder arthroscopy	80
4.1	T&O curriculum knee arthroscopy procedures mapped to expected competencies at each stage of training	104
4.2	Basic characteristics and experience of the post-CCT fellows	106
4.3	Eligibility criteria: Study B	109
4.4	Baseline characteristics of participants in Study B.	112
4.5	Usefulness: count of participant responses	119
4.6	Clarity: count of participant responses	119
4.7	Improvement: count of participant responses	120
4.8	Impact of FAST: count of participant responses	120
4.9	Impact of meniscectomy task: count of participant responses	121
4.10	Realism: count of participant responses	121
4.11	Use in training: count of participant responses	122
5.1	Selected h thresholds for each metric.	136
5.2	Difference in performance between the groups at baseline.	138
5.3	Difference in performance between groups at end of weeks 1-3. . . .	139
5.4	Attempt number where participants reached h threshold	140
5.5	Change in performance after removal of feedback.	142
6.1	Baseline characteristic of participants	156
6.2	Participant training activity during the study	164
A.1	WASP motion metrics	188
B.1	Characteristics of studies where time is used as quantitative performance metrics	197
B.2	Characteristics of studies where motion parameters are used as quantitative performance metrics	199
B.3	Characteristics of studies where global rating scales are used as quantitative performance metrics	200
B.4	Characteristics of studies where task-specific checklists are used as quantitative performance metrics	201
B.5	Characteristics of studies where error lists are used as quantitative performance metrics	202
B.6	Characteristics of studies where behavioural features are used as quantitative performance metrics	202
B.7	Characteristics of studies where force parameters are used as quantitative performance metrics	203

B.8	Characteristics of studies where single-item scales are used as quantitative performance metrics	203
B.9	Characteristics of studies where VR proprietary metrics are used as quantitative performance metrics	204
B.10	Characteristics of studies where calculated scores are used as quantitative performance metrics	204
C.1	Performance by group, measured by time taken.. . . .	213
C.2	Performance by group, measured by hand movements.	214
C.3	Performance by group, measured by minor hand movements.	215
C.4	Performance by group, measured by smoothness.	216
C.5	Performance by group, measured by maximum acceleration.	217
C.6	Performance by group, measured by stationary time.	218
C.7	Performance by group, measured by inactivity time.	219
C.8	Performance by group, measured by differences in hand dominance.	220
D.1	Probability of an alarm for the LC-CUSUM at varying h thresholds.	222
E.1	Hand movements during simulated arthroscopy list at baseline & follow-up	229
E.2	Minor hand movements during simulated arthroscopy list at baseline & follow-up	230
E.3	Smoothness during simulated arthroscopy list at baseline & follow-up	231
E.4	Time taken during simulated arthroscopy list at baseline & follow-up	232
E.5	In-theatre performance data	233

Learning on the job, particularly surgery, is relatively disorganized, it is unpredictable, it is time-consuming, and it is costly.

— Carlos A Pellegrini, former President of the American Surgical Association

1

Introduction

Contents

1.1	The practice of surgery	2
1.1.1	Ancient and pre-modern surgery	2
1.1.2	Domains of surgical skill	3
1.1.3	The role of technical skill	3
1.2	Arthroscopy	6
1.2.1	The evolution of arthroscopy	6
1.2.2	The psycho-motor challenges of arthroscopy	8
1.3	Training technical skills	9
1.3.1	Teaching technical skills	9
1.3.2	Learning technical skills	11
1.3.3	Assessing technical skills	14
1.4	Simulation	17
1.4.1	The roles of simulation in surgery	17
1.4.2	The case for simulation as a training intervention	18
1.4.3	Is surgical simulation effective?	20
1.5	Specific aims and objectives	24
1.5.1	Aim	24
1.5.2	Objectives	24

1.1 The practice of surgery

1.1.1 Ancient and pre-modern surgery

The earliest prehistoric evidence of surgical procedures is from healed trepanations in Mesolithic skulls dating from around 7000-6000 BC,¹ although it is likely that surgical interventions existed even before that but these have not been recorded or preserved.² Surgery derives from the Ancient Greek *χείρ* (cheir, meaning “hand”) and *ἔργον* (ergon, meaning “work”), and since ancient times the practice of surgery by healers was recognised as manual discipline allied to the practice of ‘physic’, or medicine, by physicians. Ancient texts and extant practice amongst primitive tribes indicate surgical interventions were predominantly limited to bone setting and the suturing or cautery of traumatic injuries,² although some surgery had a ritual³ or cosmetic⁴ purpose. Hippocrates was both a surgeon and physician, but by the Middle Ages early Christian faith healing had displaced the “barbaric”⁵ practice of surgery which was seen as distinct from the scholarly and miraculous practice of physicians and ministers.⁶ Without anaesthesia and an understanding of modern surgical pathology, surgery was necessarily quick, painful, and often brutal: a view which extended to surgeons themselves, who had become relegated to apprenticeship-trained tradesman associated with the barbers acting under the direction of university-educated physicians.⁷

Having broken away from the Company of Barber-Surgeons, the Royal College of Surgeons of England was granted its Royal Charter in 1800, the same year that Humphry Davy described his “Researches, chemical and philosophical”⁸ on nitrous oxide but did not quite appreciate the surgical implications. The advent of inhalational anaesthetic agents, along with antisepsis and haemorrhage control, triggered the golden age seen in surgery in general, and for the most part orthopaedics too.⁹ This rapidly accelerated the evolution of what had been a technical discipline subordinate to medicine into a scientific, academic, and evidence-based field driven by an understanding of surgical pathology, imaging, patient values, medical ethics, teamwork, human factors, and training over 200 years.¹⁰

1.1.2 Domains of surgical skill

Modern surgical practice requires more than just technical skill. Lord Darzi quotes FC Spencer at New York University School of Medicine saying “A skilfully performed operation is 75% decision-making and 25% dexterity”,¹¹ but even this view limits the perspective of a surgeon to an operating room technician. The formalisation of surgical training has given rise to various frameworks to describe the domains of surgical skill trainees must achieve before they qualify for independent practice (See Table 1.1). These frameworks recognise the role of the surgeon inside and outside the operating theatre in the patient care pathway, as a member of the multi-disciplinary team who care for and influence the outcomes for a patient, but also as the leader ultimately responsible for that care. Failures in intra-operative decision-making,¹² peri-operative diagnosis and practice,¹³ communication,¹⁴ management and the healthcare systems¹⁵ can all lead to patient harm.

In orthopaedics, as in other surgical subspecialties, the expanding focus on the role of non-technical domains of surgical skill^{16–18,20} and their assessment²¹ has recently displaced the predominance of technical skills which have been perceived as “easy to assess”.²² However, the technical aspects of surgery are increasingly sophisticated (e.g. image-guidance, robotic and computer-assisted surgery, arthroplasty, and arthroscopy), and as such, technical skill is undeniably a necessary quality of a good surgeon and the primary focus of this thesis.

1.1.3 The role of technical skill

Intuitively, technical skill is important in surgery: although this is not a controversial statement in itself, proving it is difficult. Technical skill is usually measured indirectly with surrogates like experience (e.g. time-in-practice and surgical volume) or complications.²³ Surgical case volume has been associated with improved outcomes: this can be interpreted as the effect of practice, but may also be influenced by selective referrals or higher volume centres attracting better staff.²⁴ Poor technical skills are implicated as a source of intra-operative error in malpractice cases¹² and increase the rate of complications in routine practice.²⁵ However, there is

Table 1.1: Domains of surgical skill required by various English-speaking professional and accrediting bodies. (* denotes the domain encompassing technical skills requirement)

Royal Australasian College of Surgeons ¹⁶	Royal College of Physicians and Surgeons of Canada ¹⁷	Accreditation Council for Graduate Medical Education ¹⁸	Intercollegiate Surgical Curriculum Programme ¹⁹
Collaboration and teamwork	Medical expert*	Practice-based learning and assessment	Applied clinical knowledge
Communication	Communicator	Patient care and procedural skills*	Applied clinical skills*
Health Advocacy	Collaborator	Systems-based practice	Professional behaviour and leadership
Judgement – clinical decision-making	Leader	Medical Knowledge	
Management and leadership	Health advocate	Interpersonal and communication skills	
Medical expertise	Scholar	Professionalism	
Professionalism and ethics	Professional		
Scholarship and teaching			
Technical skill*			

conflicting evidence about whether trainees have inferior patient-related outcomes or complications, and the effect of inexperience on complication rates may be mitigated where trainees are adequately supervised.^{26,27} For hospitals, experience affects productivity,^{28,29} and for researchers, it affects how trials of novel surgical interventions are interpreted.^{30–34} Through a number of mechanisms, technical skill is understood to influence the care we are able to provide to patients.

Beyond their direct role in patient care, technical skills are the most important professional concern for those starting surgical training.³⁵ At medical school, there is a perceived skills gap with increased emphasis placed on non-technical skills while anatomical prosection replaces dissections,³⁶ and for early years trainees there is a lack of surgical posts available³⁷ and a perception that those posts provide inadequate opportunities for technical skills preparation.^{38–41} Within early surgical training, trainees feel there is a predominantly ‘tick box’ approach to their technical skills assessment,⁴² and supervisors have insufficient time to supervise and train juniors.^{28,43} Trainees worry about technical skills decay in periods outside of training which may affect their decisions to take part in research, management or start a family,⁴⁴ and that they lack technical skills at the end of their training.^{45,46} Gender differences and confidence have also been shown to influence the way in which trainee surgeons perceive their own technical skills, with female trainees and those who are uncomfortable asking for help feeling more apprehensive about their technical skills.⁴⁷ Though it is natural that surgical trainees should wish to practice an important skill domain, technical skills also represent the ‘hidden curriculum’ by affording trainees access to the operating theatre and thus their seniors, and the patronage and career opportunities which follow.^{48,49}

Because technical skill is difficult to measure directly, studies often rely on ‘experience’ (as time-in-practice or case volume) as a surrogate marker, but non-technical skills will also have been developed with that experience. Not only are they likely to develop together, but technical skills also influence non-technical skills: higher levels of technical skill relieve attentional resources for non-technical activities intra-operatively.⁵⁰ Compared to junior trainees, more experienced trainees have

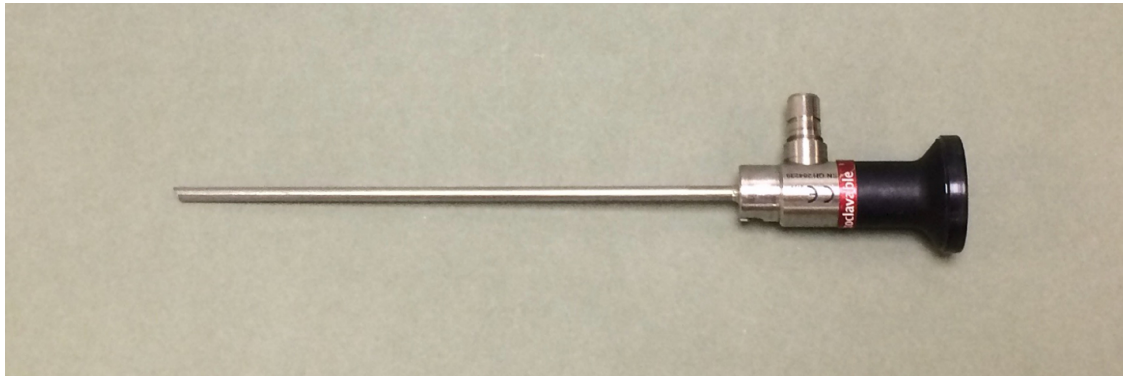
been shown to be less affected by distractions and are able to perform operative tasks more completely while engaging in both surgically relevant and irrelevant cognitive tasks.⁵¹ Outside the operating room, ‘boss technical competence’ is the strongest predictor of team job satisfaction, a vital part of the surgeon’s role as a manager and leader.⁵² As highlighted in the Canadian Medical Education Directives for Specialists (CanMEDS) framework, technical and non-technical skills are not independent, but all contribute to the ‘medical expert’.¹⁷

Technical skills have direct and indirect impacts on patient care and a lack of technical skills is specifically highlighted by trainees and trainers concerned that those in training programmes are not being prepared for independent practice.^{38,45,46} However, although there is significant assessment of non-technical skills, there is no formal assessment of technical skills at exit exams in the English-speaking world.⁵³ Meanwhile, surgery continues to evolve beyond cautery and suturing of wounds with advancing surgical technologies such as arthroscopy providing new technical skill challenges.⁵⁴

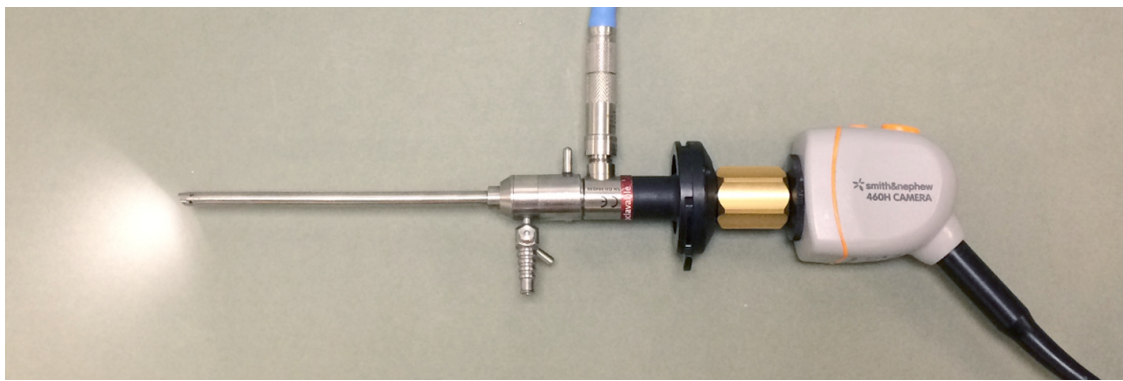
1.2 Arthroscopy

1.2.1 The evolution of arthroscopy

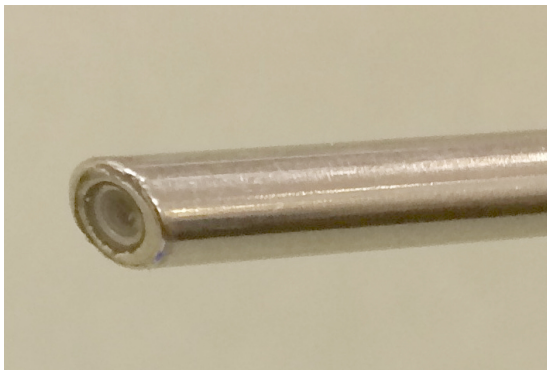
Arthroscopy is a form of minimal-access surgery where the surgeon can visualise articular spaces, and interact with intra- and peri-articular tissues.⁵⁴ Severin Nordentoft recorded the first direct arthroscopy in the knee in 1912,⁵⁵ where the surgeon views the joint directly through the aperture of the arthroscope. The rise of modern arthroscopy began with Masaki Watanabe in the 1960s and the development of fibre optics in the 1970s and 80s which allowed the arthroscopic image to be viewed indirectly (Figure 1.1).⁵⁶ Initial arthroscopies were diagnostic, but there is an expanding range of therapeutic applications.⁵⁴ Arthroscopy is becoming a more common intervention, with a 70% rise in knee arthroscopies and 750% rise in shoulder arthroscopies since 2000.^{57,58} What was once a “professor’s toy”⁵⁹ evolved into an accepted advanced technique, which has since become the commonest form of orthopaedic procedure^{60,61} and a basic competency for early years trainees.³⁸



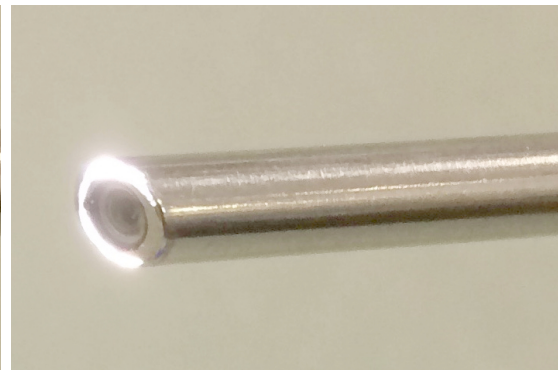
(a) An arthroscope



(b) An arthroscope with camera and cannula



(c) Fibre-optic light lead detached



(d) Fibre-optic light lead attached

Figure 1.1: Modern arthroscopic equipment. A typical 30° arthroscope (a) has a collar at its base to attach a camera and a side-port for a fibre-optic light lead. The arthroscope is used within a cannula (b) which has an irrigation tap to distend the joint with fluid. Light is introduced to the joint at the periphery of the arthroscope as in (c) & (d), then passes down the central lens to the camera.

1.2.2 The psycho-motor challenges of arthroscopy

Arthroscopy and other forms of minimal-access surgery require technical skills that pose specific challenges to trainees. Unlike open surgery, there are no common day-to-day analogues where these skills and their surrogates are developed in advance of surgical training. Challenges include:

- Axis inversion and the fulcrum effect,⁶² where hand movements move the camera tip in the opposite direction as the arthroscope pivots about the point it passes through the skin, and movements are amplified or reduced depending on the proportion of the arthroscope within the joint
- Variety of constrained, topologically complex intra-articular surgical fields,⁶³ where movement within a joint is restricted by the bony and soft tissues
- Image disruption and irrigation management,⁶⁴ where tissue, air bubbles and bleeding can impair the arthroscopic view
- Loss of haptic feedback,⁶⁵ where the ability to feel intra-articular features is diminished
- Radial distortion,⁶⁶ where images on screen are distorted at the edge of view by the nature of the size of the optical components used in arthroscopy
- Triangulation, depth perception, and reduced degrees of freedom,⁶⁷ where trainees must navigate in 3D space using 2D images

A number of technical solutions have been proposed to mitigate these effects^{68–71} such as image correction (Figure 1.2), but these are expensive elaborations to address problems that many surgeons have overcome in their training. When learning the technical skills of minimal-access surgery in the operating theatre, these challenges are of course overcome through practice but a trainee's initial attempts are associated with higher cognitive loads,⁷² slower skills acquisition,⁷³ and increased risks of harm to patients.⁶⁵ The educational benefit of observing someone else operate is most apparent when errors are made,⁶³ but these errors should

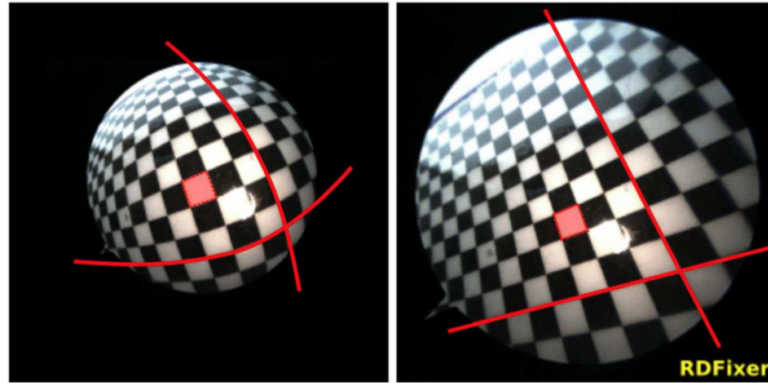


Figure 1.2: The effect of radial distortion is apparent in native (left) and corrected (right) arthroscopic views. (Reproduced from Duarte et al⁶⁸ with permission)

be conspicuously absent while watching seniors in routine arthroscopic practice. As such, the operating theatre is not the optimal environment for practising or observing arthroscopy in the initial stages of training.

Proficient arthroscopists who have ‘automatised’ the technical skills to address these challenges⁷⁴ may not realise that the operating theatre is not the best place to acquire arthroscopic skills. Modern trainees are probably not intrinsically less able than their forebears, but the pace of arthroscopic advancement means that modern trainees learn in an environment without the benefit of the intermediate technological stages their trainers were exposed to. Arthroscopic interventions have evolved in their technical complexity and become more common during the careers of current surgeons who would have come to minimal-access surgery from a position of significant surgical experience or earlier in its development, whereas modern trainees begin from a standing start in a developed discipline with fewer training opportunities.³⁹

1.3 Training technical skills

1.3.1 Teaching technical skills

Surgical training has always been provided by a form of apprenticeship. Ancient and early modern surgical training was by a traditional ‘master and apprentice’ model.² In the late 19th century, William Halsted at Johns Hopkins introduced

a model of training surgeons based on the German system, with the gradual transfer of responsibility to the trainee.⁷⁵ Still based on apprenticeship, but with a ‘pyramidal’ structure with a single professor and a number of residents in tiers below, Halsted’s model was criticized for its ‘master’ role, and in the 1930s Edward Churchill at Massachusetts General Hospital evolved it into the ‘rectangular’ structure where trainees were the responsibility of a body of trainers.⁷⁶ Despite surgical sub-specialisation and other changes to practice since then, this rectangular structure with the gradual transfer of entrustable professional activities to trainees is familiar to many surgeons today. Training was based on the principle that for the majority of trainees, service provision with a sufficient time in practice and surgical volume would lead to the transfer of the necessary skill domains, rather than on a curriculum designed for, and progress based on, a trainee’s educational needs.^{77,78} Even current trauma and orthopaedic training requires trainees to complete a minimum indicative number of index procedures before receiving their certificate of completion of training (CCT).

In recent years, fundamental changes to surgical practice and attitudes mean this model is unsustainable. Working time directives on both sides of the Atlantic have reduced the amount of time for training, and in turn, surgical volume for trainees. In the UK, to produce a consultant-led service, career pathways have been streamlined by the Calman process then Modernising Medical Careers. Combined, these have slashed the time available for training by as much as 80%, invalidating the old model.⁷⁹ Out-of-hours operating, traditionally a rich learning environment for trainees fell out of favour with the first NCEPOD report, as concurrent surgery has since.^{80,81} Trainees make theatre throughput less efficient,^{28,29} while hospitals are under extraordinary pressure to improve productivity.⁸² Trainees do not feel that the current state of affairs prepares them for independent practice.^{45,46} They are unable to achieve the guideline indicative numbers of surgical procedures to reach competence⁸³ despite these numbers being likely underestimates of the numbers needed to reach consultant-level performance.⁸⁴ Recent reforms ostensibly improve

doctors working lives, patient safety, and health service finances, but these will be short-term gains if the training does not evolve in turn.⁸¹

Competence-based training is an alternative model, focusing on what is learnt rather than time spent in training.²² The competencies required for independent practice are defined, and as trainees progress through training based on assessment and achievement of those goals, autonomy is gradually increased: with adequate supervision there is a safe, evidence-based transfer of responsibility from trainer to trainee.^{26,81} Competence-based training for technical skills relies on effective feedback and assessment, but there is evidence that current assessment methods do not promote learning,^{85,86} there is a lack of engagement from trainers,⁴² and most importantly for technical skills there is a lack of appropriate objective measures of technical skill suitable for use in theatre.^{23,87}

1.3.2 Learning technical skills

The aim of surgical training is to produce safe, independent practitioners. Modern surgical training is structured around educational principles, drawing on the psychological literature to understand the psycho-motor aspects of technical skills learning, the importance of deliberate practice, and the learning curve over time.

One of the most commonly applied models to surgical technical skill learning is the Fitts-Posner three-stage theory of motor skills acquisition.^{67,75,88} It describes an initial *cognition stage*, where a task is understood and requires active thought to perform resulting in a high cognitive load preventing adequate execution of the task and other learning,⁵⁰ followed by an *integration stage*, where practised behaviour adapts with feedback, improving task performance and reducing cognitive load. The final *automation stage* describing how tasks are performed with little cognitive input at all is often misappropriated to describe ‘expert’ performance.

The automation stage is the stated goal of trauma and orthopaedic training in the UK,⁸⁹ but this is problematic for two reasons. Firstly, *automation* is confused with *autonomy* (progress toward independent practice⁸¹) in the surgical literature.^{76,90,91} Secondly, the Fitts-Posner model describes the acquisition of everyday skills and

the plateau of performance over short periods in the general population, not the exceptional performance acquired over many years required for surgery.⁹² Although full automaticity is heralded as the feature of expert performance, Ericsson highlights it as a barrier to achieving expert performance.⁹² Automated actions have a low cognitive load as they are initiated without thought, hence they cannot be amended, and are the sign of a technician: the expert performance of a surgeon during routine surgical tasks also has low cognitive load, but remains adaptive and they “maintain cognitive control over their performance and their performance does not tend to become fully automated”.⁹²

The aphorism “practice makes perfect” is well known, and drove surgical training throughout much of the 20th century, though the nature of the practice was driven by service rather than educational needs. Ericsson’s seminal work on achieving expert performance in high-performance fields such as sport, music and medicine recognises the essential role of *deliberate practice*, especially with proximate performance feedback as a guide.⁹³ Deliberate practice differs from simple repetition by using feedback to identify areas of weakness, and monitor progress as these are systematically explored for ways to improve further. He distinguishes between the experience gained from:

Play “[unstructured] activities that have no explicit goal and that are inherently enjoyable”

Deliberate practice “[structured] activities that have been specifically designed to improve current level of performance”

Work “public performance, competitions, services rendered for pay, and other activities directly motivated by external rewards”

Routine surgical exposure in the traditional model of surgical training provides experience but this falls entirely within ‘work’: weaknesses in ability may be exposed and exhibit their consequences, but cannot be systematically explored, tested repeatedly, and refined. Modern professional standards preclude patients

being used for deliberate practice, and there is certainly no place for play in surgery. Expert performance in music and sport also rely on the “practice, rehearsal, performance” sequence, where rehearsal is the integrated practice of all skills required for performance in a work-like setting. The opportunities for rehearsal in surgery have previously been “conspicuously absent”,⁹⁴ but patient-specific 3D-printing now permits ‘dry runs’ of complex procedures,^{95,96} while pre-operative surgical simulation has been shown to provide a warm-up effect on technical performance.⁹⁷

The specific visuo-spatial and psycho-motor challenges required for arthroscopy can be overcome with experience in the ‘work’ environment, but there are generalisable components that could be isolated and practised in a simulated setting.^{98,99} For example, deliberate practice of camera navigation skills in a simulated setting is far more efficient than acquiring the same technical skills learned routine work, requiring about a third as much time as when the skills are acquired in the operating theatre.^{73,93}

Acquiring technical skills for surgery is accepted to be associated with a ‘learning curve’. Factors affecting learning curves include the type of practice,¹⁰⁰ clustering of practice,^{101,102} proximate feedback and knowledge of results,^{93,103,104} and goal setting.^{105,106} Although the learning curve is typically associated with improved performance over time, it is important to recognise that short-term rapid performance changes do not always represent sustained learning.¹⁰⁷ Innate ability will affect the start point and slope of the learning curve, and identifying innate ability has been of great interest in surgery, for example, using surrogate tasks as a means of candidate identification.¹⁰⁸ However, the literature is mixed about whether this is possible or useful. Where there are positive results relating performance to innate ability, it has not been studied whether this is a sustained benefit or an example of ‘precocious performance’ which will have little effect on the final performance level attained at the end of training.⁹² Conversely, there do appear to be some ‘untrainable’ candidates, who do not progress despite training.^{109,110} It may be valuable to detect these, but again it is unclear if this is a persistent phenomenon that does not respond to a change in approach to training.

The interaction of these factors is rarely linear and a number of mathematical models have attempted to describe learning curves,¹¹¹ but whatever their shape, the earlier stages of the learning curve are associated with more risk to patients.^{25,34,112–114} One purpose of supervised training with gradually increasing autonomy is to safely guide trainees through that period. Assessing where a trainee is on the learning curve and knowing when and how to increase autonomy is subject to potentially dangerous cognitive biases.¹¹⁵ Decreasing the autonomy of trainees is safer in the short-term, but impairs their learning and readiness for independent practice.⁸¹ An alternative approach is ‘pre-patient’ training such as simulation, which aims to “shift the learning curve” out of the operating environment.¹¹⁶

1.3.3 Assessing technical skills

Competency-based training, feedback for deliberate practice, and the analysis of learning curves all depend on a means to assess technical skill. Technical skills assessments are currently used in surgical training, in research, for trainee selection and in clinical trials. In the pre-anaesthetic age, there was only one surgical skill worth assessing: speed. Florence Nightingale noted in the Crimean War that “*cæteris paribus* the danger is in a direct ratio to the time the operation lasts”.¹¹⁷ Since the advent of anaesthesia, the length and complexity of operations has increased, but despite a recognition of the need to assess technical skill and its effect on patient outcome^{76,87} it is not formally assessed at the end of training.⁵³

A recent study showed that a majority of US orthopaedic trainers and trainees felt technical skill is not objectively measured.¹¹⁸ Although time can undoubtedly be measured objectively, care should be taken when interpreting it as a guide to performance. There are many factors beyond the operating surgeon which influence surgical time such as operating list order,¹¹⁹ surgical team composition,¹²⁰ and the presence of trainees,²⁸ notwithstanding the fact that ‘surgical time’ is often misinterpreted as ‘surgical speed’ even though ‘surgical distance’ is an unknown quantity and will vary between patients and procedures. Some surrogate objective metrics such as complication rates are fairly easy to measure, but rarity makes

them difficult to learn from and they are also prone to influences beyond the operating surgeon.^{23,24} As the Bristol heart scandal showed, complications are also statistically challenging for assessing performance.¹²¹ Surrogate metrics are at risk of creating perverse incentives for trainees, such as ‘fixation’ (leading to other areas being neglected), ‘gaming’ (where ways are found to achieve, or appear to achieve, the target performance without achieving the changes it was designed to cause) and ‘target-fatigue’.^{122,123} The IDEAL paradigm recognises the effect of technical skills on clinical trials of surgical interventions by highlighting the need to account for technical skill differences,³⁴ but provide the option to side-step the issue by performing ‘expertise-based trials’^{30,31} where surgeons are sufficiently experienced that differences in technical skill need not be measured rather than measuring and accounting for the confounding effect of experience. Unfortunately, this approach is not possible during training.

Logbooks are well established as a surrogate measure of competence⁷⁷ (see Figure 1.3). Although indicative numbers of certain surgical procedures are still used in orthopaedic training because of the link between experience and patient outcomes,^{24,114} these do not provide feedback for learning and training. Additionally, logbooks are dependent on “what comes through the door” rather than the educational needs of the trainee. There is significant disagreement over the values for indicative numbers^{124–126} and they appear to be arbitrarily set as they vary between national systems. Trainees struggle to achieve these numbers,^{83,122,123} and despite this, evidence-based attempts to quantify indicative numbers for knee and shoulder arthroscopy indicate we may significantly underestimate the volume needed to reach competence.^{84,127} Direct subjective assessments of intra-operative activity on a given case is possible using observational surveys such as global rating scales (GRS), which are complex psychometric tools validated in the research setting. Global rating scales require significant training to be applied properly in clinical practice,^{128,129} and may be improperly applied as a ‘tick box’ exercise by clinical users outside the research group.⁴² Additionally, because psychometric

tools generate an integer score, it often leads to the misinterpretation that they are objective metrics, rather than structured, quantitative, subjective metrics.¹³⁰



Figure 1.3: Royal College of Surgeons of England training record, late 19th Century. Cards were collected to log attendance at teaching.

Direct, objective measurement of intra-operative activity which depends on technical skill has proved difficult in the operating theatre.²³ Motion analysis of laparoscope movements by assessing the video has shown to be feasible, but does not appear to have been applied clinically since.¹³¹ However, in the simulated setting, there are more options. Motion analysis of surgical activity can be recorded by electromagnetic and inertial sensors, computer vision, or motion capture devices, and has been validated against other measures of technical skills such as experience and global rating scales. Not only has it been largely technologically impractical to use these methods in theatre, the direct comparison between surgeons using objective measures of surgical activity in a controlled simulated setting is very different from the operating theatre where the patients add a significant confounding variable. However, as a measure of intra-operative activity, motion analysis is well-validated,

objective, and frequently used in the simulated setting, and can provide instant feedback for independent deliberate practice.

1.4 Simulation

1.4.1 The roles of simulation in surgery

Simulation derives from the Latin *similis* meaning *like*, where the likeness in surgical simulation is to the surgical procedures, instruments, team members, and even whole operating theatres or hospital systems. Surgical simulation is a versatile modality. Clinically, surgical simulation is applied to training, assessment, candidate selection, ‘dry runs’ of complex procedures, or practice of challenging subtasks within a procedure.¹³² Non-clinical applications are also important, as simulation has a valuable public engagement role:¹³³ it can inspire and improve access to surgical careers for those who may not have considered it, and challenges the hidden curriculum in surgery. It has even been used therapeutically by patients, rather than surgeons, in the treatment of pain during medical procedures.¹³⁴ However, surgical simulation is most commonly associated with its role in training.

Simulation training is well-developed in high-stakes industries such as aviation. Flight simulators were widely available from 1927, but took-off in the 1940s as: the cost of fuel and replacing lost planes increased; the acceptance of the risk associated with flying fell with the rise of commercial air travel; and the airspeeds of planes meant fewer could be flown in a training area at any one time.¹³⁵ The combined effect of these factors was described as “reduced training space” and the parallels with surgery are clear.

Cadavers, particularly if fresh-frozen, probably represent the most accurate simulation to real surgery available today. Cadaveric specimens tend to be the preferred training modality for this reason^{136,137} and have been a well-established form of surgical training since human dissections were revived by Mondino di Luzzi at University of Bologna in the 14th century.¹³⁸ Some animal models have similar anatomy to human, e.g. the porcine knee,¹³⁹ with the advantage of being easier to source and, outside the UK, simulation is permitted using live anaesthetised animals

e.g. for practising damage control surgery for catastrophic injuries. However, both human and animal cadaveric simulations have significant cost and licensing barriers,¹⁴⁰ as well as ethical implications.^{141,142} Modern simulation techniques range from the high-tech, high-fidelity ‘realistic’ models encompassing many surgical skills and procedures in physical or virtual-reality space with the ability to provide feedback to the user and collect data on performance, to the simple, low-fidelity setting which may only address a few generalizable components of surgery. For example, a recently described low fidelity arthroscopy model includes a grapefruit with a nut hidden inside.¹⁴³ In arthroscopy, numerous joint types have been simulated in virtual and physical space, and combined with measures of technical skill have demonstrated the ability to differentiate between participants with different levels of experience and show learning on simulators with training. Some studies have even shown that simulated training subjectively affects intra-operative behaviour.^{144,145} Given the variety of simulation options, what place do they have in training, and how do we gauge their effectiveness?

1.4.2 The case for simulation as a training intervention

Surgical simulation may be able to address some of the issues raised about technical skills training in orthopaedics today. Brown notes the pedagogical, ethical and practical case for simulation in surgery.⁹⁹ It provides a safe, risk-free environment for pre-patient training during the cognition stage of learning, and deliberate practice to support operating theatre experience in the integration stage where competencies and learning curves can be explored. Objective evaluation of simulated performance can guide trainers about how and when they safely cede autonomy to their trainees.^{81,115} These satisfy the pedagogical and ethical case, being an educationally learner centred environment and reducing the need to practice on patients.^{99,116}

In practical terms, efficiency of training is also increased. In theatre, practice on patients is educationally and ethically dubious,¹⁴⁶ and reduces theatre efficiency.^{28,29} In the skills lab, trainees can reach the same performance levels for camera navigation more quickly in than if they rely on organically occurring opportunities in theatre,⁷³

and can deliberately practice on the novel aspects of a procedure alone, like ACL graft preparation.¹³² Regular distributed simulation of rarer procedures can smooth out the dependence on “what comes through the door” to maintain skills acquired in ‘boot camps’ or on specialist courses. This is an important benefit, as many of the challenges arise from the ‘reduced training space’ effect of working time directive, stream-lined career paths, and service demands which affect training opportunities and contact time with one’s trainer. More efficient training allows more time for service, or at least reduces the tension between training and service provision in posts.

There is a tension in modern surgical practice: arthroscopic and minimally-invasive surgery has been recognised by the Royal College of Surgeons of England ‘Commission on the Future of Surgery’,¹⁴⁷ yet the evidence behind the common ‘training’ arthroscopic procedures such as diagnostic arthroscopy,¹⁴⁸ meniscectomy,¹⁴⁹ and subacromial decompression¹⁵⁰ has been shown to be lacking. In the future, these procedures and the training they provide may become less common. The advantages of minimal-access surgery such as improved recovery times and reduced scarring are worth pursuing, but without common training procedures to develop the required skills trainees will need another environment to learn arthroscopic and minimal-access surgical techniques.

When implementing surgical simulation, it is important to consider the appetite for it, how it would integrate with current training, and the likely impact. Although it is popular with patients¹³³ and trainees,¹⁵¹ and promoted by educational bodies,¹⁵² practical details on how it is to be implemented and how it will be funded are vague: two recent studies have shown that their simulation training programmes caused unanticipated rises in demands on faculty time and raised the cost of training by 15 times, without objective evidence of improved training outcomes.^{153,154} Simulation should not top-up what is lost to working time restrictions, but is part of an entirely new approach to training.⁷⁸ Simulation provides the potential for self-directed study, like keeping up-to-date with journals. Arthroscopy may be particularly suited to simulation training, given its psycho-motor challenges and its technological features that make it more ‘simulatable’ than open surgery. Given the potential costs,

there is also a moral case to ensure simulation actually works.¹⁵⁵ Even the most high-fidelity simulations will never replace ‘real’ in-theatre operating experience.¹⁵⁶ Poorly designed studies are difficult to interpret clinically,¹⁰⁸ can reinforced bad habits,¹⁵⁷ and time spent in simulation detracts from operative experience and service provision.¹⁵⁸ However, the most significant criticism is that despite its potential costs, there is currently limited evidence it actually alters intra-operative behaviour.

1.4.3 Is surgical simulation effective?

Validity

The term validity comes from the Latin *valere*, meaning *strong*. Although the educational, ethical, and practical case for simulation may be compelling to some, our decisions about the implementation of surgical simulation need to be strong: not based on the strength of conviction, but the strength of evidence. This validation does not investigate a quality a simulator intrinsically has or has not, as validity is a “property of the inferences, not instruments, [so] validity must be established for each intended interpretation”.¹⁵⁹ In the context of surgical simulation for technical skills training, this depends on both the simulator and the performance metrics, which together form the construct: varying either or both of these, or the setting in which the construct is applied, will affect the validity of our interpretations. The current framework of validity is a unitary concept with multiple sources evidence to support each required interpretation of the construct as a measure of what it claims to be measuring. Although the discrete model of ‘types’ of validity has fallen out of favour, these can be useful to provide sources of evidence for unitary interpretation of validity:¹⁶⁰

Construct validity The degree to which a test measures what it claims to be measuring. In technical skills training, this is often interpreted as the ability of a construct to distinguish between participants based on their level of technical skills. This is the only form of validity in the contemporary framework

Face validity An assessment of a simulated construct's subjectively reported 'realism' compared to the task being simulated

Content validity An assessment of the features within a construct, and to what degree the relevant elements of the task being simulated are present

Concurrent validity An assessment of how two constructs agree in the evidence they supply in support of the unitary validity. The constructs may differ in the content or form of the task and the performance metrics, depending on the interpretations to be made

Reliability and accuracy

When using surgical simulation to train or assess technical skill, the metrics chosen to measure performance are intrinsically linked to the interpretations made about the validity of simulation. As with any form of measurement, it is essential to understand how reliable and how accurate the measure is at representing the real world in order to make decisions about how to act on the results:

Reliability An assessment of how reproducible results from a construct are when tested repeatedly

Accuracy An assessment of how close results from a construct represent the real or true result. This can be challenging to measure in complex, multi-faceted domains such as technical skill

The reliability and accuracy of a construct are independent concepts, and both can be measured statistically to help understand the validity of conclusions made about a construct.

The Kirkpatrick model

Beyond the properties of the construct, it is important to understand how these have been effectively applied to the setting being simulated. Kirkpatrick describes a model for evaluating the impact of training sessions on the trainee and their work¹⁶¹

in the industrial setting which has been co-opted to surgery.^{162,163} In the context of simulation training for technical skills, the model distinguishes the trainees *reaction* and engagement with a training programme (Level 1) and any *learning* within the training environment only (i.e. changes in simulated performance; Level 2), from the *behaviour* change in the real-world (i.e. improved technical skills in theatre; Level 3) and the *results* for the task being simulated (i.e. patient outcomes improvements; Level 4). This model has parallels to the phases used in clinical trials (see Table 1.2). It is a criticism of surgical simulation that Levels 1 and 2 of the Kirkpatrick model are often self-evident, but the real-world effect, often called the *transfer validity* or *transfer of training* (from the skills lab to the operating room) has not been adequately demonstrated.

How valid is surgical simulation?

There is a place for high and low fidelity simulations at different stages of surgical training, but the strength of the evidence which establishes how simulation training affects performance relies on the quality of the metrics used.¹⁶⁴ Metrics may be selected because they are easily measured, but are often inappropriate, subjective or prone to bias. Surrogates may be recorded for what one actually wants to measure. Poorly chosen metrics lead to perverse incentives in training, poor measures of performance, and lack of engagement with the training process.²²

The reason simulation works in aviation is because the same valid, accurate, objective performance metrics are used in the simulated and real-world setting. This makes performance directly comparable between the two and allows pilots-in-training to calculate a ‘training effectiveness ratio’, where the costs of simulation to reach a required level of airborne performance correlate with a reproducible ratio to savings in flying time, fuel spent, and *in extremis* loss of aircraft and pilots.¹³⁵ In contrast, there is currently limited evidence in arthroscopy to support the notion that surgical simulation training leads to objective improvements in intra-operative performance. The strongest data supporting surgical simulation are based on subjective assessments of intra-operative performance^{144,145} using global

Table 1.2: Comparison of the Kirkpatrick levels with clinical trial phases.

Kirkpatrick Model	Description	Description	Clinical trial phase
1 - Reaction	The degree to which participants find the training favourable, engaging and relevant to their jobs	Adverse effects, first in man	1
2 - Learning	The degree to which participants acquire the intended knowledge, skills, attitude, confidence and commitment based on their participation in the training	'Proof of concept' detecting any clinical effect or biological activity	2a
3 - Behaviour	The degree to which participants apply what they learned during training when they are back on the job	Dose optimisation for biological effect that is clinically relevant	2b
4 - Results	The degree to which targeted outcomes occur as a result of the training and the support and accountability package	Assesses the effectiveness of the therapeutic dose on patient outcomes	3
		Post marketing surveillance	4

rating scales and task-specific checklists. These tools place an intense requirement on faculty¹⁵⁴ which is costly¹⁵³ and they do not permit informed, independent practice. Although objective methods of technical skills assessment do exist, these have been limited to the skills lab so assessment of transfer validity to the operating theatre and behavioural change leading to improved patient outcomes hasn't yet been measured. The benefits of surgical simulation on outcomes are even more difficult to measure, as surgical outcomes are not always proximate to the surgery itself or attributable to any single member of the surgical team. To date there is no real evidence demonstrating clinical and cost-effectiveness of simulation training in surgery, and so no dedicated funding streams exist to develop simulation training or simulation skills centres in the UK.

Current arthroscopic simulation has received some validation, but lack of objective in-theatre metrics means there is a knowledge gap about whether simulation training objectively improves intra-operative technical skills.

1.5 Specific aims and objectives

1.5.1 Aim

The aim of this thesis is to investigate whether technical skills practise through arthroscopic simulation leads to objective improvements in intra-operative arthroscopic surgery performance.

This will be achieved through the following objectives.

1.5.2 Objectives

1. Identify how quantitative metrics for technical skill assessments have been applied to arthroscopy
2. Validate suitable objective, quantitative metrics of skills assessment for use in both a simulated and surgical setting
3. Develop and validate a simulated arthroscopy list for deliberate practice of the fundamental technical skills associated with arthroscopic surgery

4. Investigate the impact of simulated arthroscopy, with performance feedback using objective metrics, on the acquisition of arthroscopic skills
5. Demonstrate the objective impact of simulation training by the transfer of acquired arthroscopic skills to the operating theatre

*There is no excuse for the for the surgeon to learn on
the patient.*

— William J. Mayo

2

Systematic review of quantitative methods of assessing arthroscopic performance

Contents

2.1	Introduction	27
2.2	Aims	28
2.3	Methods	28
2.3.1	Study identification	28
2.3.2	Registration	31
2.4	Results and analysis	32
2.4.1	Description of included studies	32
2.4.2	Quantitative methods of assessing arthroscopic performance	35
2.4.3	Time	35
2.4.4	Motion parameters	42
2.4.5	Global rating scales	46
2.4.6	Task-specific checklists	50
2.4.7	Error lists	53
2.4.8	Behavioural features	55
2.4.9	Force parameters	57
2.4.10	Single-item scales	59
2.4.11	Virtual reality proprietary metrics	61
2.4.12	Calculated scores	63
2.4.13	Fluoroscopy	66
2.4.14	Correlation of quantitative measures of arthroscopic per- formance to intra-operative behaviour	67
2.4.15	Correlation of quantitative measures of arthroscopic per- formance to patient-reported outcome measures	69

2.5 Discussion	69
2.5.1 Summary of findings	69
2.5.2 Strength and limitations	69
2.5.3 In the context of the wider literature	70
2.5.4 Implications of this study	71

2.1 Introduction

Technical performance is recognised as a fundamental domain of surgical expertise,^{16,165,166} and is accepted to improve with experience and play a role in patient outcomes.^{25,27,33,167} Measuring technical performance should play a role in talent identification and selection,¹¹⁰ progress through training programmes,^{103,168} and adjusting outcomes of clinical trials,³⁰ but is hampered because the measurement of technical performance in and out of the operating theatre is challenging.^{11,87,169,170} This may explain why technical performance is not formally assessed as part of orthopaedic surgery exit examinations in the United Kingdom, United States or Australia.^{166,171,172} although the examinations consider technical skills concepts in the questions within the written and viva assessments, there are no practical examinations to assess technical performance.

High-quality surgical training requires reliable measures of technical performance that can be used by trainees, trainers, and training programmes.¹⁷³ Traditional assessment methods are subjective and prone to cognitive bias.^{11,174} Metrics which are used for feedback on performance need to be validated to show they are relevant and important measures of performance which show improvement with the acquisition of technical skills. The performance metrics chosen must be delivered cost-effectively in a training or assessment situation and provide useful guidance to trainees about how to improve. Ultimately, performance metrics should be correlated to patient-reported outcome measures (PROMs).¹⁰³

All surgical interventions require technical expertise, but arthroscopy has specific psycho-motor challenges that have no day-to-day analogues a junior trainee can draw upon. This hampers insight into self-assessment of technical skills

performance which is typically impaired at the earlier stages of training and with more complex procedures,¹⁷⁵ with potential consequences for patient safety.^{121,176} Accurate technical performance metrics proximate to the event are essential to provide the feedback needed for deliberate practice and the acquisition of expert performance.¹⁰⁰

2.2 Aims

Investigate what quantitative metrics have been used to assess arthroscopic performance, and how they correlate to patient outcomes.

2.3 Methods

2.3.1 Study identification

A systematic review of the literature was conducted according to preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines, with the study aims, design and search strategy specified using the participants, interventions, comparators, outcomes and study design (PICO) process:¹⁷⁷

Participants All arthroscopic practitioners, from zero previous experience to fully-qualified independent practitioners

Interventions Orthopaedic arthroscopic surgery of knee, shoulder, hip, ankle, elbow and wrist joints, with surgery taking place in a theatre or simulated setting

Comparators Any or no comparison groups

Outcomes Any proximate quantitative measure of technical performance

Study design All designs

A comprehensive search of the EMBASE and MEDLINE databases via Ovid and the MEDLINE database via PubMed was conducted to week 4 of July 2016,

using appropriate equivalent search terms for the database subheading mappings (see Appendix B.1).

The search results were imported to *Covidence*,¹⁷⁸ the Cochrane Collaboration’s online systematic review management platform, which performed automatic de-duplication. Two reviewers independently screened the remaining titles and abstracts, assisted by highlighting of keywords imported to *Covidence*(Table 2.1). Screened studies underwent full text review in duplicate and were selected for inclusion in this systematic review according to the inclusion and exclusion criteria in Table 2.2. Exclusion reasons were recorded and studies were only excluded where there was mutual agreement by both reviewers. References from the remaining articles were screened for studies which had been missed by the search strategy. From the articles selected for inclusion, data were extracted about the:

- study design, purpose and OCEBM level of evidence
- the joint studied, the setting for the tasks, and nature of the environment
- participant population and demographics
- performance metrics recorded, how they were applied, and Kirkpatrick level demonstrated
- a summary of study conclusions

Table 2.1: Keywords to assist in title and abstract screening on *Covidence* platform. (* is a ‘wild-card’ character)

Keyword			
performance	time	sentinel	proficien*
global	metric	error	competen*
scale	ASSET	valid	learning
score	IGARS	construct	haptic
path	OSATS	Delphi	
distance	BAKS	skill	

Reviews were included as a source for missed references. Narrative synthesis was planned for the study conclusions, with descriptive analysis of the number

Table 2.2: Inclusion and exclusion criteria.

Inclusion	Exclusion
Original research or review	Letters, editorials, and lectures/courses
Any assessment of orthopaedic arthroscopic ability using a quantitative method	Non-orthopaedic and non-arthroscopic procedures
Any level of surgeon (student, SHO, registrar, consultant, attending)	Assessing performance by patient outcomes only
Surgical procedures	Comparisons of different procedures
Simulated procedures - bench top, VR, animal, cadaveric	Assessment of departments or institutions rather than individual practitioners
Surgeon performance	Descriptions and instructions/assessments of surgical techniques
Orthopaedic arthroscopic surgery of knee, shoulder, hip, ankle, elbow, and wrist joints	Only non-surgeon performance recorded e.g. anaesthetic, scrub staff, radiology
Surgery taking place in a theatre or simulated setting (including cadaveric, bench top, virtual reality, and animal models)	Technical documents on simulator design, development studies, experiences, and publications chiefly about product development rather than deployment and use
Any method of assessing patient outcome (PROMS, complications, pain visual analogue scale, pain diary, return to activities e.g. work, sport)	No proximate performance metrics

of quantitative assessment methods available, how many are in common use, and their validity and training efficacy that have been demonstrated by the included studies. It was not anticipated that meta-analysis would be possible due to expected heterogeneous outcomes and study settings.

The Kirkpatrick levels were developed to monitor the effect of training interventions.¹⁶¹ Not all of the studies included would use quantitative metrics to assess the impact of a specified training programme, so the Kirkpatrick levels would not apply to all studies. Due to studies with non-training interventions, the Kirkpatrick levels were used as a framework to describe how the quantitative performance metrics had been applied:

Reported where a quantitative metric is reported but not associated with technical performance in a training (simulation) or work (intra-operative) setting

Learning where a quantitative metric is associated with performance differences in a training (simulation) setting. Equivalent to Kirkpatrick level 2, if after a training intervention

Behaviour where a quantitative metric is associated with performance differences in a work (intra-operative) setting. Equivalent to Kirkpatrick level 3, if after a training intervention

Outcomes where a quantitative metric is associated differences to patient outcome. Equivalent to Kirkpatrick level 4, if after a training intervention

2.3.2 Registration

The study was registered with the Prospero database (CRD42016046584).

Changes to registered protocol

Two additional forms of surgery were included that were not anticipated by the initial protocol. The search strategy identified papers investigating these endoscopic spinal surgery and suprascapular nerve decompression where ‘arthroscopic’ techniques were in use.^{179,180} This was not a deliberate omission, but came about as they are relatively rare in the UK and did not come up during discussion of initial screening criteria. While it could be argued these are not strictly arthroscopic, but rather endoscopic or minimal-access procedures, an executive decision was made by the author to include such papers as they are orthopaedic in nature and could reasonably occur in the T&O training, and present a transferable skill-set for arthroscopy proper. In support of this decision, it is worth noting hip arthroscopy also includes a significant extra-articular component.¹⁸¹

The original protocol planned that all selected studies would be assessed for quality using the MERSQI score, an instrument to assess the quality of education research.¹⁸² Not only was the search strategy likely to include some non-educational studies, but once the protocol was submitted, it was noted that “quality scales and resulting scores are not an appropriate way to appraise clinical trials”.¹⁸³ as a result, the suggested the Cochrane Risk of Bias tool would be used instead. Not all retrieved studies would be assessed for bias, but only clinical trials where the

technical performance metrics identified were likely to affect decisions about real-surgical practice on patients (studies showing changes to Behaviour and Outcomes; Kirkpatrick level 3 and 4).

2.4 Results and analysis

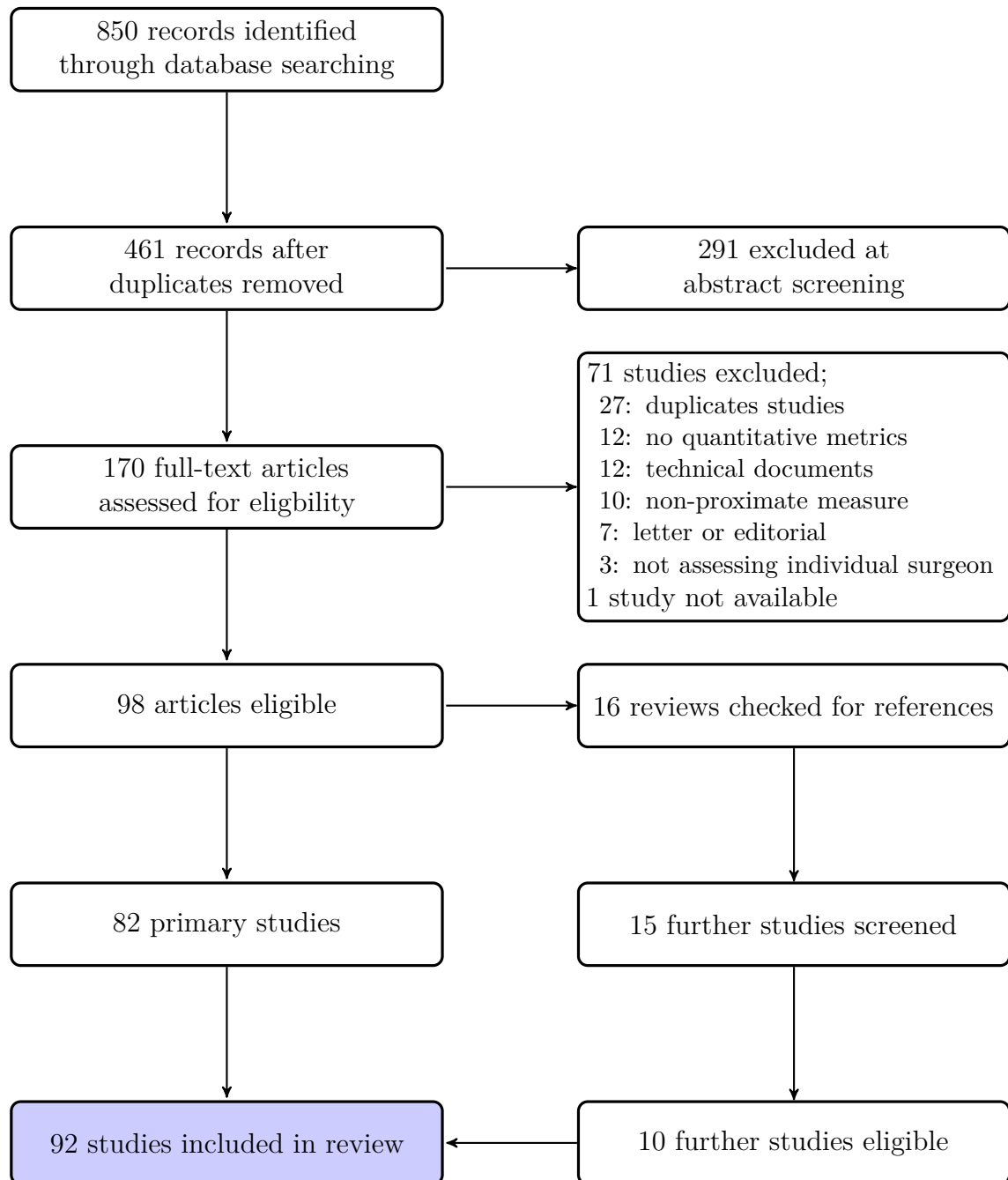
The search of EMBASE and MEDLINE via Ovid, and MEDLINE via PubMed, yielded 361, 247, and 242 articles respectively. After de-duplication, 461 unique articles were screened by both reviewers: 278 were excluded outright by both reviewers, 150 were put forward for full text review, and disagreements about the remaining 33 articles were resolved by discussion with 20 going forward to full text review. There was good agreement between the reviewers at this stage (Table 2.3a), with linear-weighted kappa 0.90 (95% CI 0.86-0.94). Of the 170 articles submitted to full-text review, 98 were indentified as eligible (82 primary, 16 reviews), 71 were excluded (Table 2.3b). One could not be retrieved for full-text review. The 82 primary studies underwent data extraction, and the 16 reviews' reference lists were used to identify 15 further articles for screening, of which 10 were included for data extraction, bringing the total to 92 included studies (Figure 2.1).

Table 2.3: Inter-rater agreement between Reviewers A & B.

(a) at initial screening					(b) at full-text review			
B					B			
		Exclude	Maybe	Include			Exclude	Include
A	Exclude	278	4	0	A	Exclude	71	0
	Maybe	20	0	0		Include	0	98
	Include	9	0	150				

2.4.1 Description of included studies

Ninety-two studies reported quantitative technical performance metrics in arthroscopy, in 17 randomised controlled trials, 63 cohort studies and 12 case-series. As per Thomsen et al,¹⁶² quantitative metrics of arthroscopic performance were being validated in 47 studies, efficacy of skills training was being assessed in 42 studies,

**Figure 2.1:** PRISMA flow diagram

and three studies had another primary purpose: these studies used simulators to assess the effect of 3D viewing systems,¹⁸⁰ telementoring,¹⁸⁴ and distraction⁵¹ on surgeon performance. Purely descriptive studies were excluded, but 22 of the studies described the development of a quantitative metric and validated it in the same study. Kirkpatrick levels were recorded in about a third (32/92; 35%) of studies which assessed a formal training intervention using quantitative metrics: 28 studies showed that the training programme led to learning in the simulated environment (Kirkpatrick level 2), and four studies showed that the training programme led to intra-operative behavioural change (Kirkpatrick level 3). No studies assessed reaction to the metrics alone (Kirkpatrick level 1), and no studies showed intra-operative behaviour change led to improvements in patient outcomes (Kirkpatrick level 4). A total of 2514 participants had their performance assessed in the included studies. There were 631 participants with no arthroscopic experience (typically medical students), 1163 trainee surgeons, 442 experienced surgeons who had completed their training, and the experience of 264 participants was unclear from the studies. There were 10 single-surgeon series, 21 studies assessing participants as a single group, 26 studies with two groups, and 35 studies with three or more groups. For an overview of the characteristics of the included studies, see Table 2.4.

The majority (72/92; 78%) of studies assessed performance in the simulated setting alone, compared to 16 in the operating theatre. Four studies assessed performance in both the simulated and real-life surgical setting. These were randomised controlled trials investigating the transfer of simulation training into improvements in intra-operative performance.^{144,145,185,186} Where simulators were used, virtual-reality was commonest environment with 38 studies, then 25 anatomic bench-top model and 7 generic bench-top model studies, 16 cadaver studies, two animal studies and one desktop, mouse-operated computer simulation. Knee arthroscopy was the commonest procedure to be studied (46 studies), followed by shoulder arthroscopy (38), while hip (3), ankle (1), spine (2), elbow (1), and wrist (1) were less frequent. There were also three studies which compared the cross-over of laparoscopic skills to arthroscopy.^{187–189} Diagnostic arthroscopic tasks

were most commonly assessed (55 studies), then therapeutic tasks (29 studies) and part-tasks (17 studies.)

2.4.2 Quantitative methods of assessing arthroscopic performance

A wide variety of quantitative performance metrics were reported, and many studies reported more than one metric. Quantitative metrics reported in the studies were grouped where they were judged to be thematically linked to aid discussion where metrics have similar properties, even though the techniques used to acquire the results may not be directly interchangeable between metrics within a group. Table 2.5 shows how each class of metric has been applied to the included studies, using Kirkpatrick levels as a framework for both training and non-training scenarios, as described above (pg. 30).

2.4.3 Time

This groups quantitative metrics where the time taken to perform a procedure is recorded. It is calculated by a range of techniques, including using a stopwatch, times from theatre logs, or the internal clock on computer or VR simulators.

Time is the commonest quantitative metric to be recorded in the retrieved studies, being reported in 72 studies. In 49 studies a difference was shown with either increasing experience or with training in a simulated environment, and 11 studies demonstrated an improvement with experience or training in the operating theatre. PROMs were recorded in six studies, but time taken did not correlated with patient outcomes. Table 2.6 summarises the evidence demonstrated for the application of time as a quantitative performance metric and overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 197).

Time is an intuitively understandable quantitative measure of performance for surgeons and lay people alike, and requires no sophisticated equipment to record. Lister famously observed that “Anaesthetics have abolished the need for operative speed and allow time for careful procedure”.¹¹ Thankfully, time is less pressing

Table 2.4: Characteristics of included studies.

		Number of studies
Primary studies identified		92
Study design	Randomised control trials	17
	Cohort	63
	Case series	12
Purpose	Descriptive	22
	Validation	47
	Efficacy	42
	Other	3
Kirkpatrick level	1	0
	2	28
	3	4
	4	0
	Non-training intervention	60
Task type	Diagnostic	55
	Therapeutic	29
	Part-task	17
Groups	Single surgeon	10
	One group	21
	Two groups	26
	Three or more groups	35
Setting	Simulator only	72
	Simulator and theatre	4
	Theatre only	16
Environment	Virtual reality	38
	Benchtop - anatomic	24
	Benchtop - generic	7
	Cadaver	17
	Animal	2
	Computer	1
Joint studied	Knee	46
	Shoulder	38
	Hip	3
	Spine	2
	Ankle	1
	Elbow	1
	Wrist	1
	Laparoscopic skills crossover	3
	Generic	7
Number of participants	Total	2514
	No arthroscopic experience	631
	Trainee surgeons	1163
	Experienced surgeons	442
	Unclear	278

Table 2.5: Quantitative metrics applied to the assessment of arthroscopic technical skill identified in 92 included studies.

	Reported	Learning	Behaviour	Outcomes
Time	72	49	11	0
Motion parameters	35	30	0	0
Global rating scales	28	22	4	0
Task-specific checklist	25	10	2	0
Error lists	4	3	0	0
Behavioural features	15	7	0	0
Force parameters	11	4	1	0
Single-item scale	6	3	1	0
VR proprietary metrics	6	2	0	0
Calculated score	12	6	0	0

today than it had been earlier in Lister’s career, though time still is important as lengthy procedures increase risks of complications^{190,191} and trainees take longer to perform procedures which increases costs and decreases theatre utilisation.^{28,29} Time is influenced by case mix and the theatre team, which are beyond the surgeon’s control.¹¹¹ In athletics or motor sports, time is used as surrogate measure for speed, where distance may not be known but is at least standardised. By contrast, ‘surgical distance’ is unknown, and even within a single procedure type there is patient variation which prevents the direct comparison between time taken and ‘speed’ of the surgeon. In the operating theatre, the experience of the team and the surgeon both affect the length of the case.²³ A “task may be done quickly, by not necessarily well”,¹⁶⁴ but operative time is related to theatre throughput, and thus remuneration, and the easy availability of time gives it undue prominence as a measure of performance. Time is easily understood and can be ‘manipulated’ by the operating surgeon, so it should be used cautiously as a performance metric to avoid perverse incentives.

Table 2.6: Evidence demonstrated for application of time as a quantitative performance metric.

Paper	Application	KP	Summary of findings
Akhtar ¹⁹² 2013 Arthroscopy	Learning	2	Time taken and path length performance showed statistically significant improvement with practice, and could distinguish participants based on experience.

KP: Kirkpatrick level demonstrated in study overall; NT: Non-training intervention. Continued on next page

Table 2.6: Evidence demonstrated for application of time as a quantitative performance metric.

Paper	Application	KP	Summary of findings
Akhtar ¹⁹³ 2013 Arthroscopy	Learning	NT	Time taken, path length and roughness performance improve with simulator practice, but statistically significant performance decay after 1 year practice hiatus.
Akhtar ¹⁸⁹ 2016 J Surg Educ	Learning	2	Transferable skills between arthroscopy and laparoscopy. Time taken, path length and hand speed performance in trained groups showed statistically significant improvement relative to controls.
Alvand ¹¹⁰ 2011 JBJS Br	Learning	2	Trained participants time taken, path length and hand movements performance is statistically significantly better than untrained during shoulder arthroscopy. Some participants fail to reach 'competence' despite training.
Alvand ¹⁰⁹ 2011 JBJS Am	Learning	2	Statistically significant difference in time taken, path length and hand movements performance between participants successful and unsuccessful at achieving task competence with practice. Some participants may lack innate ability for arthroscopy.
Alvand ¹⁹⁴ 2012 JBJS Am	Reported	NT	Good correlations of performance by simple visual parameters with GRS & motion analysis metrics
Alvand ¹⁹⁵ 2013 Arthroscopy	Learning	2	Learning curves of time taken, path length and hand movements performance shows statistically significant improvement over study, with moderate correlation to GRS performance.
Anakwenze ¹⁹⁶ 2013 JSES	Behaviour	NT	Time taken showed statistically significant improvement over series, but no correlation to PROMS
Andersen ¹⁹⁷ 2011 Acta Orthop	Learning	2	Time taken, path length, and collision depths statistically significantly improve with training.
Angelo ¹⁹⁸ 2015 Arthroscopy	Reported	NT	Time taken, errors and sentinel errors made showed statistically significant differences with operative experience, but no difference in the surgical steps
Angelo ¹²⁹ 2015 Arthroscopy	Reported	2	Surgical steps, errors and sentinel errors made statistically significantly improved with proficiency-based training, with no differences in time taken
Angelo ¹⁹⁹ 2015 Arthroscopy	Learning	NT	Time taken and errors made showed statistically significant differences with operative experience, but no difference in the surgical steps or 'sentinel' errors
Bayona ²⁰⁰ 2008 J Robot Surg	Reported	2	Reports face and content validity of simulator. Score is described, with contribution of each metric weighted by expert opinions of its importance, but not used to assess performance
Bliss ²⁰¹ 2005 SHTI	Reported	2	Landmark identification, time taken and tissue collisions showed statistically significant improvement over training period. 'Minimal' degradation/skills decay over 4 weeks, but no statistics reported
Braman ²⁰² 2015 Arthroscopy	Learning	NT	Time taken, accidental contacts and dropped pegs show statistically significant improvement with operative experience on some simulated tasks
Butler ²⁰³ 2013 JBJS Am	Learning	2	Practice on dry models before cadaveric simulation statistically significantly reduced the number of attempts and time taken required to reach a minimum GRS performance standard.
Camp ²⁰⁴ 2016 JBJS Am	Learning	2	Cadaveric training group showed a statistically significant improvement in time taken and GRS performance after training, not seen in other groups.
Cannon ¹⁴⁵ 2014 JBJS Am	Reported	3	TSCL shows statistically significant difference between experimental and control groups in theatre, but no difference in GRS or time taken

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.6: Evidence demonstrated for application of time as a quantitative performance metric.

Paper	Application	KP	Summary of findings
Cannon ²⁰⁵ 2014 JBJS Am	Learning	NT	Time taken shows statistically significant improvement with operative experience, but no differences in summary score.
Castricini ²⁰⁶ 2013 Musc Surg	Behaviour	NT	Time taken showed statistically significant improvement over series, but no correlation to PROMS
Chang ²⁰⁷ 2016 CORR	Learning	NT	Time taken, GRS and global score performance show statistically significant improvement with operative experience. 'Care for soft tissues' results not reported.
Cowan ⁵¹ 2016 Orthopedics	Learning	NT	All trainees susceptible to distraction, but senior trainees still require less time, cause fewer injuries, and look at hand less often
Cunningham ²⁰⁸ 2016 KSSTA	Behaviour	NT	Time taken showed statistically significant improvement over series. No conversions to open after 10th patient.
Del Piñal ²⁰⁹ 2012 J Hand Surg Am	Behaviour	NT	Time per procedure decreased over series, with no conversions to open
Dunn ¹⁸⁶ 2015 J Surg Educ	Behaviour	3	Time taken and GRS performance improves in both control and experimental groups, but no differences between the two groups. Task specific checklist and single item scale performance show no change.
Elliot ²¹⁰ 2012 Arthroscopy	Learning	NT	Time taken, TSCL and ASAF-Knee scoring metric show statistically significant improvement with operative experience
Escoto ²¹¹ 2012 SHTI	Learning	NT	Time taken and average force applied improved with operative experience
Escoto ²¹² 2013 IEEE	Learning	2	Apparent improvement in time taken and path length performance after practice, but no statistical confirmation. Device able to measure forces, but not used or reported.
Fucentese ²¹³ 2015 KSSTA	Learning	NT	Time taken and path length show statistically significant improvement with operative experience, but no difference in camera horizontal alignment
Gandhi ²¹⁴ 2015 Arthroscopy	Reported	NT	Positive correlations between performance on internet-based tool and VR simulator path length and time taken performance
Gandhi ²¹⁵ 2015 SECEC	Reported	NT	No consistent statistically significant correlation between participants subjective scoring of their manual dexterity and hand-eye coordination with time taken, path length, roughness or VR global score performance.
Gilmer ²¹⁶ 2015 Arthroscopy	Learning	2	Knot-tying speed and confidence improve with training, but no change in knot strength
Gomoll ²¹⁷ 2007 AJSM	Learning	NT	Statistically significant improvements in time taken, collisions, path length and probe velocity with operative experience
Gomoll ²¹⁸ 2008 AJSM	Learning	2	Time take, collisions, velocity and path length statistically significantly improved after 3 years of clinical training, equivalent performance to matched historical group.
Guttman ¹²⁶ 2005 Arthroscopy	Behaviour	NT	Time taken performance shows statistically significant decrease over series. No PROMS recorded.

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.6: Evidence demonstrated for application of time as a quantitative performance metric.

Paper	Application	KP	Summary of findings
Henn ²¹⁹ 2013 Arthroscopy	Learning	2	Time taken and GRS performance statistically significantly improve between baseline and follow up in experimental and control groups, but but no inter-group differences in GRS performance
Howells ²²⁰ 2008 Arthroscopy	Learning	NT	Time taken, path length and hand movements performance show statistically significant performance differences between surgeons and non-surgeons
Howells ¹⁴⁴ 2008 JBJS Br	Learning	3	Time taken and motion parameters show learning curve in skills lab. In theatre, GRS and TSCL show statistically significant performance differences for simulation trained group compared to controls
Howells ²²¹ 2009 JBJS Am	Learning	2	Learning curve shows time taken, path length and hand movements performance statistically significantly improves with practice, with significant skills decay after 6 month hiatus
Jacobsen ⁹¹ 2015 JBJS Am	Learning	NT	Time taken, path length and roughness performance distinguishes participants based on experience. Calculated score failed 1/13 qualified surgeons and passed 2/13 novices in high stakes assessment.
Kim ²²² 2010 KSSTA	Behaviour	NT	Time taken performance shows statistically significant improvement at over initial cases in series, then plateau. No improvement in PROMS or patient post-op range of motion over series.
Kirby ²²³ 2015 JBJS Am	Learning	NT	Arthroscopic performance score from machine learning technique more accurately defines the groups when sensors worn at elbow. Correlates with GRS performance. Elbow and wrist worn sensors performance show strong correlation.
Koehler ²²⁴ 2015 Arthroscopy	Learning	NT	More experienced surgeons had statistically significantly better GRS performance during knee and shoulder arthroscopy in theatre. Safety domain of GRS has poor inter-rater agreement (ICC 0.30)
Konan ²²⁵ 2011 JBJS Am	Behaviour	NT	Time taken performance statistically significantly improves over series. PROMs collected but not analysed
Lee ²²⁶ 2013 KSSTA	Behaviour	NT	Time taken showed statistically significant improvement over series, but no correlation to PROMS
Martin ²²⁷ 2011 JBJS Am	Learning	2	Time taken performance shows correlation between cadaveric and VR simulator models.
Martin ²²⁸ 2012 JBJS Am	Learning	NT	Fair correlations between time taken and path length performance to PGY and previous arthroscopic experience
Martin ²²⁹ 2015 Foot Ankle Int	Learning	2	ASSET total score, safety score and TSCL show statistically significant differences between intervention and control group on anatomic and cadaveric models
McCarthy ²³⁰ 2006 CORR	Learning	NT	Time taken, arthroscope path length and number of loose bodies identified showed statistically significant improvement with operative experience.
Middleton ²³¹ 2016 JBJS Am	Learning	NT	Performance recorded by each GRS distinguishes between participants based on experience, shows strong correlation of performance measured with other GRS and with time/motion metrics
Nicandri ²³² 2015 OJSM	Learning	2	Time taken, path length, smoothness and probe pressure performance statistically significantly improved with practice on the simulator. No significant change in 'completeness of procedure' VR metric.
Pedowitz ²³³ 2002 Arthroscopy	Learning	NT	Time taken and path ratio improve with operative experience, but no change in force parameters

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.6: Evidence demonstrated for application of time as a quantitative performance metric.

Paper	Application	KP	Summary of findings
Pollard ²³⁴ 2012 JBJS Am	Learning	2	Time taken, path length and hand movement performance show statistically significant improvement with practice. The lateral position demonstrated a more challenging learning curve than supine position.
Ponce ¹⁸⁴ 2014 JBJS Am	Reported	NT	The use of telementoring equipment did not statistically significantly affect time take performance during arthroscopic rotator cuff repair or stabilisation.
Price ⁸⁴ 2015 JBJS Br	Learning	NT	GRS performance improves with experience, and correlates to time taken, path length and hand movements. Approx 170 arthroscopies estimated to achieve consultant level GRS performance.
Rahm ²³⁵ 2016 BMC Surg	Learning	2	Overall score, calculated from time taken and path length, improves with practice over first 2 hours of training, then plateaus
Rebolledo ²³⁶ 2015 AJSM	Learning	2	Simulator training improved time taken and IGI performance in shoulder model, but not knee model
Rose ²³⁷ 2015 Arthroscopy	Learning	NT	Inconsistent support for better performance with operative experience across three tasks with multiple metrics.
Safir ¹⁸⁸ 2012 SHTI	Learning	2	Time taken performance is statistically significantly improved during simulated arthroscopy in participants with laparoscopic practice.
Smith ²³⁸ 1999 SHTI	Learning	NT	Trained orthopaedic surgeons perform a more complete examination, take less time and make fewer collisions than medical students and other surgeons across a range of exercises
Srivastava ²³⁹ 2004 JSES	Learning	NT	Time taken and 'surgical navigation score' statistically significantly improve with operative experience.
Strom ¹⁸⁷ 2004 Surg Endosc	Reported	2	No effect of visuospatial training on time taken, path length, collisions or overall score during virtual arthroscopy noted
Stunt ²⁴⁰ 2014 KSSTA	Learning	NT	Time taken performance distinguishes between participants based on experience
Stunt ²⁴¹ 2014 KSSTA	Learning	NT	Time taken performance distinguishes between participants based on experience
Tashiro ²⁴² 2009 CORR	Learning	NT	Time taken, path length, velocity, and peak & average force performance show statistically significant improvement with operative experience
Thomsen ¹⁸⁰ 2004 Arthroscopy	Reported	NT	No significant difference in the time taken or TSCL performance between the 2D and 3D display systems
Tuijthof ²⁴³ 2010 KSSTA	Learning	NT	Time taken performance distinguishes between participants based on experience
Tuijthof ²⁴⁴ 2011 CORR	Learning	NT	Time taken performance distinguishes between participants based on experience during initial trials, but not at later attempts
Wang ²⁴⁵ 2011 Spine J	Behaviour	NT	Time taken showed statistically significant improvement over series, but no correlation to PROMS, blood loss or hospital stay.

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.6: Evidence demonstrated for application of time as a quantitative performance metric.

Paper	Application	KP	Summary of findings
Waterman ¹⁸⁵ 2016 Orthopedics	Learning	3	In theatre, statistically significant difference GRS safety domain between groups, but not time, total GRS, or TSCL . On simulator, simulation group statistically significantly improved time and path length performance compared to controls.
Wong ²⁴⁶ 2015 KSSTA	Learning	NT	Experienced surgeons complete statistically significantly more knots in the time allowed. No change in performance at 6 months.
Yamakado ¹⁷⁹ 2015 Arthroscopy	Behaviour	NT	Time taken performance statistically significantly improves over series. No link to PROMS or complications.

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention

2.4.4 Motion parameters

This groups quantitative metrics where position, velocity, acceleration or jerk is recorded during a procedure, which can be used to calculate path-length, the number of movements made or the jerkiness/smoothness of a procedure. Lower-order derivatives of position can be differentiated with respect to time to calculate higher-order derivatives (e.g. velocity to acceleration), but the reverse, though mathematically possible, incorporates errors due to the loss of constants.

Thirty-five studies were identified which assessed motion parameters as a measure of performance. Thirty of these showed an effect of learning or experience in the simulated setting, but none have been used in the operating theatre. The commonest recorded metric was path length, the total distance travelled by the instruments or hands of the user. To record the motion parameters, virtual reality or electromagnetic means were used in all but one study: both these techniques are able to provide accurate positional information in a simulated setting (which is used to calculate motion parameters), but these technologies do not translate easily into the operating theatre. This might explain why, despite extensive validation in the simulated setting, motion parameters have not been used to assess intra-operative behaviour or patient outcomes. Table 2.7 summarises the evidence demonstrated for the application of motion parameters as quantitative performance metrics and

overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 199).

Motion parameters are believed to represent performance because with increasing experience, it is expected fewer unnecessary movements are required to complete a procedure. As they are recorded automatically, not by an observer, they are truly objective, quantitative measures of performance. However, as with interpreting time taken, ‘surgical distance’ is not knowable on a procedure by procedure basis. Due to inter-patient differences it is difficult to draw conclusions based on motion parameters alone about how a surgeon performed, in contrast with global rating scales which can subjectively account for the complexity of the case and patient-specific factors.

Table 2.7: Evidence demonstrated for application of motion parameters as quantitative performance metrics.

Paper	Motion parameter [method - device]	Application	KP	Summary of findings
Akhtar ¹⁹² 2013 Arthroscopy	Path length of arthroscope, path length of probe, roughness [VR - ArthroMentor]	Learning	2	Time taken and path length showed statistically significant improvement with practice. Distinguishes participants based on experience.
Akhtar ¹⁸⁹ 2016 J Surg Educ	Path length of left hand, path length of right hand, speed of left hand, speed of right hand [VR - ArthroMentor]	Learning	2	Transferable skills between arthroscopy and laparoscopy. Time taken, path length and hand speed in trained group showed statistically significant improvement relative to controls.
Akhtar ¹⁹³ 2013 Arthroscopy	Path length of arthroscope, path length of probe, roughness [VR - ArthroMentor]	Learning	NT	Time taken, path length and roughness performance improve with simulator practice, but statistically significant performance decay after 1 year practice hiatus.
Alvand ¹¹⁰ 2011 JBJS Br	Path length total; total number of hand movements [EM - Polhemus Patriot]	Learning	2	Trained participants time taken, path length and hand movements performance is statistically significantly better than untrained during shoulder arthroscopy. Some participants fail to reach ‘competence’ despite training.
Alvand ¹⁰⁹ 2011 JBJS Am	Path length total; total number of hand movements [EM - Polhemus Patriot]	Learning	2	Statistically significant difference in time taken, path length and hand movements performance between participants successful and unsuccessful at achieving task competence with practice. Some participants may lack innate ability for arthroscopy.
Alvand ¹⁹⁵ 2013 Arthroscopy	Path length total; total number of hand movements [EM - Polhemus Patriot]	Learning	2	Learning curves of time taken, path length and hand movements performance shows statistically significant improvement over study, with moderate correlation to GRS performance.
Alvand ¹⁹⁴ 2012 JBJS Am	Path length total; total number of hand movements [EM - Polhemus Patriot]	Reported	NT	Good correlations of performance by simple visual parameters with GRS & motion analysis metrics

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.7: Evidence demonstrated for application of motion parameters as quantitative performance metrics.

Paper	Motion parameter [method - device]	Application	KP	Summary of findings
Andersen ¹⁹⁷ 2011 Acta Orthop	Path length of arthroscope; path length of probe [VR - insightMIST]	Learning	2	Time taken, path length, and collision depths statistically significantly improve with training.
Chang ²⁰⁷ 2016 CORR	Path length of arthroscope, path length of instruments, roughness [VR - insight ArthroVR]	Learning	NT	Time taken, GRS and global score performance show statistically significant improvement with operative experience. 'Care for soft tissues' results not reported.
Escoto ²¹² 2013 IEEE	Path length of left and right hand and instruments [EM - Aurora/ISCAD]	Learning	2	Apparent improvement in time taken and path length performance after practice, but no statistical confirmation. Device able to measure forces, but results not reported.
Fucentese ²¹³ 2015 KSSTA	Path length of arthroscope, path length of probe [VR - ArthroS]	Learning	NT	Time taken and path length show statistically significant improvement with operative experience, but no difference in camera horizontal alignment
Gandhi ²¹⁴ 2015 Arthroscopy	Path length of arthroscope, path length of instruments, roughness [VR - insightArthroVR]; Actual user track distance [ASATs]	Reported	NT	Positive correlations between performance on internet-based tool and VR simulator path length and time taken performance
Gandhi ²¹⁵ 2015 SECEC	Path length of arthroscope, path length of instruments, roughness [VR - insight ArthroVR]	Reported	NT	No consistent statistically significant correlation between participants subjective scoring of their manual dexterity and hand-eye coordination with time taken, path length, roughness or VR global score performance.
Gomol ²¹⁷ 1 2007 AJSM	Path length of arthroscope, path length of probe, speed of probe [VR - ProCedicus]	Learning	NT	Statistically significant improvements in time taken, collisions, path length and probe velocity with operative experience
Gomoll ²¹⁸ 2008 AJSM	Path length of arthroscope, path length of probe, speed of probe [VR - ProCedicus]	Learning	2	Time take, collisions, velocity and path length statistically significantly improved after 3 years of clinical training, equivalent performance to matched historical group.
Howells ²²⁰ 2008 Arthroscopy	Path length total, number of hand movements total [EM - Polhemeus Patriot]	Learning	NT	Time taken, path length and hand movements performance show statistically significant performance differences between surgeons and non-surgeons
Howells ²²¹ 2009 JBJS Am	Path length total, number of hand movements total [EM - Polhemeus Patriot]	Learning	2	Learning curve shows time taken, path length and hand movements performance statistically significantly improves with practice, with significant skills decay after 6 month hiatus
Howells ¹⁴⁴ 2008 JBJS Br	Path length total, number of hand movements total [EM - Polhemeus Patriot]	Learning	3	Time taken and motion parameters show learning curve in skills lab. In theatre, GRS and TSCL show statistically significant performance differences for simulation trained group compared to controls

KP: Kirkpatrick level demonstrated in study overall; NT: Non-training intervention. Continued on next page

Table 2.7: Evidence demonstrated for application of motion parameters as quantitative performance metrics.

Paper	Motion parameter [method - device]	Application	KP	Summary of findings
Jackson ²⁴⁷ 2012 JBJS Am	Path length total, number of hand movements total [EM - Polhemeus Patriot]	Learning	2	Learning curve shows time taken, path length and hand movements performance statistically significantly improves with practice, without skills decay after 6 month hiatus
Jacobsen ⁹¹ 2015 JBJS Am	Path length and roughness of arthroscope and probe [VR - ARTHRO Mentor]	Learning	NT	Time taken, path length and roughness performance distinguishes participants based on experience. Calculated score with the thresholds described would fail 1 of 13 qualified surgeons and pass 2 of 13 novices in high stakes assessment.
Kirby ²²³ 2015 JBJS Am	Path length[VR - ArthroS]; Smoothness, Hand movements, Hand acceleration, Dominance, Stationary time [Inertial - WASP]	Learning	NT	Arthroscopic performance score from machine learning technique more accurately defines the three experience groups when sensors worn at elbow, and correlates well with GRS performance. Elbow and wrist worn sensors recorded hand movements and smoothness performance show strong correlation.
Martin ²²⁸ 2012 JBJS Am	Path length of arthroscope, path length of probe [VR - ArthroVR]	Learning	NT	Fair correlations between time taken and path length performance to PGY and previous arthroscopic experience
McCarthy ²³⁰ 2006 CORR	Path length of arthroscope, path length of probe [VR - SKATS]	Learning	NT	Time taken, arthroscope path length and number of loose bodies identified showed statistically significant improvement with operative experience.
Middleton ²³¹ 2016 JBJS Am	Path length total [VR - ArthroS]	Learning	NT	Performance recorded by each GRS distinguishes between participants based on experience, shows strong correlation of performance measured with other GRS and with time/motion metrics
Nicandri ²³² 2015 OJSM	Path length of arthroscope, path length of probe, smoothness [VR - TolTech]	Learning	2	Time taken, path length, smoothness and probe pressure performance statistically significantly improved with practice on the simulator. No significant change in 'completeness of procedure' VR metric.
Pedowitz ²³³ 2002 Arthroscopy	Path length ratio (%age of optimal) [VR - Procedicus]	Learning	NT	Time taken and path ratio improve with operative experience, but no change in force parameters
Pollard ²³⁴ 2012 JBJS Am	Path length total, number of hand movements total [EM - Polhemeus Patriot]	Learning	2	Time taken, path length and hand movement performance show statistically significant improvement with practice. The lateral position demonstrated a more challenging learning curve than supine position.
Price ⁸⁴ 2015 JBJS Br	Path length total, number of hand movements total [EM - Polhemeus Patriot]	Learning	NT	GRS performance improves with experience, and correlates to time taken, path length and hand movements. Approx 170 arthroscopies estimated to achieve consultant level GRS performance.
Rahm ²³⁵ 2016 BMC Surg	Path length of arthroscope, path length of probe, path length total [VR - ArthroS]	Learning	2	Overall score, calculated from time taken and path length, improves with practice over first 2 hours of training, then plateaus

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.7: Evidence demonstrated for application of motion parameters as quantitative performance metrics.

Paper	Motion parameter [method - device]	Application	KP	Summary of findings
Rose ²³⁷ 2015 Arthroscopy	Path length of left and right hands, difference in path lengths [VR - ArthroVision]	Learning	NT	Inconsistent support for better performance with operative experience across three tasks with multiple metrics.
Smith ²³⁸ 1999 SHTI	Path length ratio (%age of optimal) [VR - ProCedicus]	Reported	NT	Trained orthopaedic surgeons perform a more complete examination, take less time and make fewer collisions than medical students and other surgeons across a range of exercises
Srivastava ²³⁹ 2004 JSES	Path length ratio [VR - ProCedicus]	Learning	NT	Time taken and 'surgical navigation score' statistically significantly improve with operative experience.
Strom ¹⁸⁷ 2004 Surg Endosc	Path length ratio [VR - ProCedicus]	Reported	2	No effect of visuospatial training on time taken, path length, collisions or overall score during virtual arthroscopy noted
Tashiro ²⁴² 2009 CORR	Path length and speed of arthroscope and probe [EM - Aurora]	Learning	NT	Time taken, path length, velocity, and peak & average force performance show statistically significant improvement with operative experience
Waterman ¹⁸⁵ 2016 Orthopedics	Path length of arthroscope and probe [VR - ArthroVR]	Learning	3	In theatre, statistically significant difference GRS safety domain between simulation and control group, but not time, total GRS, or TSCL performance. On simulator, simulation group statistically significantly improved time and path length performance compared to controls.

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention

2.4.5 Global rating scales

This groups quantitative metrics where performance is assessed by an observer, who uses a Likert scale to rate the performance of the surgeon in a structured, subjective manner across the important skill domains of a procedure. When designing a global rating scale (GRS), experts in the field decide on the relevant domains, and also provide anchor statements for the Likert scales in each domain to guide the observer. Anchor statements are often mapped to the Dreyfus model.^{103,248}

Twenty-eight studies were identified, 22 of which showed the GRS assessing performance differences in a simulated setting, and 5 showed the GRS assessing a behavioural change in the operating theatre. The commonest GRS were Arthroscopic Surgery Skill Evaluation Tool (ASSET;²⁴⁹ 11 studies) and Basic Arthroscopic Knee Skill Scoring System (BAKSSS;²⁵⁰ 9 studies) which were both designed for knee

arthroscopy. In the simulated setting, GRS have been used extensively to show their construct validity and the effect of training on simulated performance (Kirkpatrick level 2). The significant amount of positive validation work published may be due to the fact that many of these scales are often modified versions of one-another.^{195,203,251} Middleton et al²³¹ showed three common global rating scales performed equally well at accurately distinguishing participants based on their ability across six simulated exercises. Table 2.9 summarises the evidence demonstrated for application of global rating scales as quantitative performance metrics and overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 200).

Global rating scales are common, with both generic and procedure-specific varieties. Some have been extensively validated, they can be used in the operating theatre, and in non-orthopaedic fields they have even been used to link performance to patient outcomes,²⁵ so would seem a reasonable choice for use in the operating theatre to demonstrate transfer of training and the effect on patient outcomes in orthopaedics. However, although they are understood to be a quantitative assessment of performance and also permit assessment of quality, both of these beliefs are based on logical fallacies.

Firstly, the summed scores of scale items produce a single integer. This is understood to be an objective measure (the names of some scales even include the word ‘objective’^{144,252}), but Likert scales are inherently subjective, relying on observers judgements of the performance relative to their interpretations of the anchor statements. That subjectivity is exacerbated where the GRS design leaves blank junctions between anchor statements (typically at 2 and 4), or includes ambiguous and “double-barrelling” statements²⁵³ (Table 2.8). The quantitative results encourage certain statistical analysis which are controversial, because although the data are ordinal, the ‘distance’ between anchor statements is unknown.¹³⁰ Secondly, the scales are modelled on the Dreyfus performance characteristics, with the scale-length matched to the five ‘Novice-to-Expert’ levels and the anchor statements written to mirror these. This creates a circular reasoning behind the scoring of expertise in global rating scales, which is also subjective. It limits the ability of

Table 2.8: Global rating scale anchor statement examples which require subjective interpretation

Feature	Explanation	Example	Reference
Ambiguous anchor statement	Adjectives with no specific meaning require judgement of assessor	“Frequently used unnecessary force on tissue...”	Reznick 1997 ²⁵⁴
Pseudo-quantitative statement	Without a measurement tool, apparently quantitative statements are estimates	“Graceful and dextrous throughout procedure with camera always centred and correctly oriented”	Koehler 2013 ²⁴⁹
Double-barrelling	Interpreting responses where there are two variables in one statement is ambiguous	“[1]Constantly overshoots target. [2]Slow to correct.”	Bayona 2014 ²⁵⁵
Begging the question	Statement presupposes the claim	“Expertly uses both hands in complementary manner to provide optimum performance.”	Insel 2009 ²⁵⁰

these scales to rank participants drawn from certain proficiency level since the instrument is designed to group them into that level, not differentiate within it.

Table 2.9: Evidence demonstrated for application of global rating scales as a quantitative performance metric.

Paper	Global rating scale	Application	KP	Summary of findings
Alvand ¹⁹⁵ 2013 Arthroscopy	BAKSSS	Learning	2	Learning curves of time taken, path length and hand movements performance shows statistically significant improvement over study, with moderate correlation to GRS performance.
Alvand ¹⁹⁴ 2012 JBJS Am	BAKSSS	Reported	NT	Good correlations of performance by simple visual parameters with GRS & motion analysis metrics
Bayona ²⁵⁵ 2014 JBJS Am	IGARS	Learning	NT	Novel GRS with high internal consistency and inter-rater agreement, which distinguishes between participants based on experience
Butler ²⁰³ 2013 JBJS Am	BAKSSS	Learning	2	Practice on dry models before cadaveric simulation statistically significantly reduced the number of attempts and time taken required to reach a minimum GRS performance standard.
Camp ²⁰⁴ 2016 JBJS Am	ASSET	Learning	2	Cadaveric training group showed a statistically significant improvement in time taken and GRS performance after training, not seen in other groups.
Cannon ¹⁴⁵ 2014 JBJS Am	GRAP	Reported	3	TSCL shows statistically significant difference between experimental and control groups in theatre, but no difference in GRS or time taken
Chang ²⁰⁷ 2016 CORR	Modified GRS	Learning	NT	Time taken, GRS and global score performance show statistically significant improvement with operative experience. ‘Care for soft tissues’ results not reported.
Dunn ¹⁸⁶ 2015 J Surg Educ	ASSET	Behaviour	3	Time taken and GRS performance improves in both control and experimental groups, but no differences between the two groups. Task specific checklist and single item scale performance show no change.
Dwyer ²⁵⁶ 2015 AJSM	ASSET	Learning	NT	GRS and TSCL performance shows statistically significant improvements with time in training, with some ceiling effects noted

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.9: Evidence demonstrated for application of global rating scales as a quantitative performance metric.

Paper	Global rating scale	Application	KP	Summary of findings
Dwyer ²⁵⁷ 2016 Arthroscopy	ASSET	Learning	NT	GRS and TSCL performance showed statistically significant improvement with time in training on each task
Henn ²¹⁹ 2013 Arthroscopy	GOALS	Learning	2	Time taken and GRS performance statistically significantly improve between baseline and follow up in experimental and control groups, but but no inter-group differences in GRS performance
Hodgins ²⁵⁸ 2014 Arthroscopy	BAKSSS	Learning	NT	Learning curve shows 40% of trainees met TSCL criteria for competence, and 5% of trainees met GRS criteria for competence over training period
Howells ¹⁴⁴ 2008 JBJS Br	OSATS	Behaviour	3	Time taken and motion parameters show learning curve in skills lab. In theatre, GRS and TSCL show statistically significant performance differences for simulation trained group compared to controls
Hoyle ²⁵¹ 2011 Should Elb	GRSSA	Behaviour	NT	GRS scores have good internal consistency, but poor inter-rater reliability
Insel ²⁵⁰ 2009 JBJS Am	BAKSSS	Learning	NT	GRS shows statistically significant improvement with number of procedures and postgraduate year, while TSCL improvement shows early ceiling effect
Kirby ²²³ 2015 JBJS Am	BAKSSS	Learning	NT	Arthroscopic performance score from unsupervised machine learning technique more accurately defines the three experience groups when sensors worn at elbow, and correlates well with GRS. Elbow and wrist worn sensors recorded hand movements and smoothness show strong correlation.
Koehler ²⁴⁹ 2013 AJSM	ASSET	Learning	NT	GRS score correlates with level of training. TSCL results not reported.
Koehler ²²⁴ 2015 Arthroscopy	ASSET	Behaviour	NT	More experienced surgeons had statistically significantly better GRS performance during knee and shoulder arthroscopy in theatre. Safety domain of GRS has poor inter-rater agreement (ICC 0.30)
Koehler ²⁵⁹ 2013 JBJS Am	ASSET	Learning	NT	GRS performance more likely to be awarded 'passing' score with increasing time in training and case load.
Martin ²²⁹ 2015 Foot Ankle Int	ASSET	Learning	2	ASSET total score, safety score and TSCL show statistically significant differences between intervention and control group on anatomic and cadaveric models
Martin ¹³⁹ 2015 CORR	OAAS	Learning	NT	No OAAS performance difference between human and porcine cadaveric knees. Good correlation between years of experience and OAAS score
Middleton ²³¹ 2016 JBJS Am	Multiple	Learning	NT	Performance recorded by each GRS distinguishes between participants based on experience, shows strong correlation of performance measured with other GRS and with time/motion metrics

KP: Kirkpatrick level demonstrated in study overall; NT: Non-training intervention. Continued on next page

Table 2.9: Evidence demonstrated for application of global rating scales as a quantitative performance metric.

Paper	Global rating scale	Application	KP	Summary of findings
Nicandri ²³² 2015 OJSM	ASSET	Learning	2	Time taken, path length, smoothness and probe pressure performance statistically significantly improved with practice on the simulator. No significant change in ‘completeness of procedure’ VR metric.
Olson ²⁶⁰ 2013 CORR	BAKSSS	Learning	NT	GRS performance statistically significantly higher for senior cf. junior residents. Fair inter-rater reliability for total GRS score (0.69) and its use to define pass/fail (0.54).
Price ⁸⁴ 2015 JBJS Br	BAKSSS	Learning	NT	GRS performance improves with experience, and correlates to time taken, path length and hand movements. Approx 170 arthroscopies estimated to achieve consultant level GRS performance.
Slade Shantz ²⁵² 2013 Arthroscopy	OAAS	Learning	NT	GRS and single-item VAS show statistically significant improvement with operative experience
Waterman ¹⁸⁵ 2016 Orthopedics	ASSET	Reported	3	In theatre, statistically significant difference GRS safety subdomain between simulation and control group, but not time, total GRS, or TSCL performance. On simulator, simulation group statistically significantly time and path length performance than controls.
Wong ²⁴⁶ 2013 KSSTA	GRS	Reported	2	Self-calculated ‘proficiency score’ correlates well with observer-recorded GRS and TSCL

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention

2.4.6 Task-specific checklists

This groups quantitative metrics where performance is assessed by the presence of positive actions recorded by an observer, typically identifying and interacting with specific anatomical structures or displaying desirable behaviours, representing the completeness of the procedure performed. Experts in the field define the essential steps, a task-specific checklist (TSCL), for a procedure.

Twenty-five studies were identified using task-specific checklists to assess performance. Of these, 14 studies used published instruments (Table 2.10) and the remainder used locally described checklists. Two studies by Cannon et al¹⁴⁵ and Howells et al¹⁴⁴ demonstrated intra-operative behavioural change as a result of simulation training using these metrics (Kirkpatrick level 3). The TSCL identified were constructed by a mixture of formal Delphi panel, expert discussions, instructions from industry operative technique manuals, reviewing of textbooks and literature,

Table 2.10: Published task-specific checklists identified in the included studies

Task-specific checklist	Studies	Construction
Copernicus	4	Delphi process
Arthroscopic Surgery Skill Evaluation Tool (ASSET)	2	Delphi process
Basic Arthroscopic Knee Skill Scoring System (BAKSSS)	2	Literature review
Arthroscopic Skills Assessment Form (ASAF)	2	Expert review
American Academy of Orthopaedic Surgeons (AAOS)	2	Undocumented
Orthopaedic Competence Assessment Project (OCAP)	1	Undocumented
Objective Practical Assessment Tool (OPAT)	1	Delphi process

or by an undocumented process. Table 2.11 summarises the evidence demonstrated for the application of TSCL as quantitative performance metrics and overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 201).

Task-specific checklists are important deconstructions of the essential elements of a procedure which become increasingly objective with “unambiguous step definition”.²⁶¹ However, they do not assess the manner in which steps are performed, so minimal experience is required to perform well on these scales and they have early ceiling effects impairing their use as a performance measure.^{250,258,262} Task-specific checklists may supplement a global rating scale to standardise how a procedure is performed,^{249,250} or add a procedure-specific element to an otherwise generic scale,²⁵⁴ but in practice global rating scales are often used without a task specific checklist due to the ceiling effects.^{203,231,258,260} Some task specific checklists try to avoid the ceiling effects by grading each step on the list with subjective quality assessments¹⁴⁵ or pseudo-quantitative estimations of the completeness of a step,²⁶³ losing the dichotomous objectivity which is their main advantage over global rating scales. Task-specific checklists are common quantitative measures of arthroscopic performance which have been used in theatre but performance has not been linked to PROMs. By design they represent the minimum standard for performing a procedure,²⁴⁹ with marked ceiling effects as a result.

Table 2.11: Evidence demonstrated for application of task-specific checklists as quantitative performance metrics.

Paper	Checklist	Application	KP	Summary of findings
Angelo ¹⁹⁸ 2015 Arthroscopy	Copernicus	Report	NT	Time taken, errors and sentinel errors made showed statistically significant differences with operative experience, but no difference in the surgical steps
Angelo ¹²⁹ 2015 Arthroscopy	Copernicus	Learning	2	Surgical steps, errors and sentinel errors made statistically significantly improved with proficiency-based training, with no differences in time taken
Angelo ¹⁹⁹ 2015 Arthroscopy	Copernicus	Report	NT	Time taken and errors made showed statistically significant differences with operative experience, but no difference in the surgical steps or 'sentinel' errors
Angelo ¹²⁸ 2015 Arthroscopy	Copernicus	Reported	NT	Face and content validation of error and TSCL metrics
Bliss ²⁰¹ 2005 SHTI	Local	Report	2	Landmark identification, time taken and tissue collisions showed statistically significant improvement over training period. 'Minimal' degradation/skills decay over 4 weeks, but no statistics reported
Cannon ¹⁴⁵ 2014 JBJS Am	Local	Behaviour	3	TSCL shows statistically significant difference between experimental and control groups in theatre, but no difference in GRS or time taken
Cowan ⁵¹ 2016 Orthopedics	ASAF	Reported	NT	All trainees susceptible to distraction, but senior trainees still require less time, cause fewer injuries, and look at hand less often
Dunn ¹⁸⁶ 2015 J Surg Educ	Local	Reported	3	Time taken and GRS performance improves in both control and experimental groups, but no differences between the two groups. Task specific checklist and single item scale performance show no change.
Dwyer ²⁵⁶ 2015 AJSM	Local	Learning	NT	GRS and TSCL performance shows statistically significant improvements with time in training, with some ceiling effects noted
Dwyer ²⁵⁷ 2016 Arthroscopy	Local	Learning	NT	GRS and TSCL performance showed statistically significant improvement with time in training on each task
Elliot ²¹⁰ 2012 Arthroscopy	ASAF	Learning	NT	Time taken, TSCL and ASAF-Knee scoring metric show statistically significant improvement with operative experience
Hodgins ²⁵⁸ 2014 Arthroscopy	BAKSSS	Learning	NT	Learning curve shows 40% of trainees met TSCL criteria for competence, and 5% of trainees met GRS criteria for competence over training period
Howells ¹⁴⁴ 2008 JBJS Br	OCAP	Behaviour	3	Time taken and motion parameters show learning curve in skills lab. In theatre, GRS and TSCL show statistically significant performance differences for simulation trained group compared to controls
Insel ²⁵⁰ 2009 JBJS Am	BAKSSS	Learning	NT	GRS shows statistically significant improvement with number of procedures and postgraduate year, while TSCL improvement shows early ceiling effect
Koehler ²⁴⁹ 2013 AJSM	ASSET	Reported	NT	ASSET score correlates with level of training. TSCL results not reported.
Martin ²²⁹ 2015 Foot Ankle Int	Local	Learning	2	ASSET total score, safety score and TSCL show statistically significant differences between intervention and control group on anatomic and cadaveric models

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.11: Evidence demonstrated for application of task-specific checklists as quantitative performance metrics.

Paper	Checklist	Application	KP	Summary of findings
Martin ¹³⁹ 2015 CORR	AAOS	Learning	NT	No OAAS performance difference between human and porcine cadaveric knees. Good correlation between years of experience and OAAS score
McCarthy ²³⁰ 2006 CORR	Local	Reported	NT	Time taken, arthroscope path length and number of loose bodies identified showed statistically significant improvement with operative experience.
Slade Shantz ²⁵² 2013 Arthroscopy	AAOS	Reported	NT	GRS and single-item VAS show statistically significant improvement with operative experience
Smith ²³⁸ 1999 SHTI	Local	Learning	NT	Trained orthopaedic surgeons perform a more complete examination, take less time and make fewer collisions than medical students and other surgeons across a range of exercises
Srivastava ²³⁹ 2004 JSES	Local	Reported	NT	Time taken and surgical navigation score statistically significantly improve with operative experience.
Talbot ²⁶³ 2015 Arthroscopy	OPAT	Reported	NT	Strong positive correlation between novel TSCL performance and usual assessment method. Fair inter-rater reliability (0.60), with excellent intra-rater reliability(0.82).
Thomsen ¹⁸⁰ 2004 Arthroscopy	Local	Reported	NT	No significant difference in the time taken or TSCL performance between the 2D and 3D display systems
Waterman ¹⁸⁵ 2016 Orthopedics	ASSET	Reported	3	In theatre, statistically significant difference GRS safety subdomain between simulation and control group, but not time, total GRS, or TSCL performance. On simulator, simulation group statistically significantly time and path length performance than controls.
Wong ²⁴⁶ 2013 KSSTA	Local	Reported	2	Self-calculated ‘proficiency score’ correlates well with observer-recorded GRS and TSCL

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention

2.4.7 Error lists

This groups quantitative metrics where performance is assessed by the presence of negative actions or deviations from usual procedure that may cause harm, defined by experts in the field.

Only one error list was identified, used in four studies for assessing Bankart repairs by the group that developed it. This list distinguishes between 77 errors (“deviations from optimal performance”) and 20 sentinel errors (“...serious deviations from optimal performance [that jeopardise] the success/desired result of the procedure or create iatrogenic insult...”).¹²⁸ This group have demonstrated that error lists can be used to assess learning in a simulated setting, but not applied

the instrument in the operating theatre. Table 2.12 summarises the evidence demonstrated for the application of error lists as quantitative performance metrics and overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 202).

Error-observation is an effective form of training,⁶³ and error lists could be useful to guide trainees to learn from the mistakes of others and increase awareness of potential difficulties, reducing complacency and allow trainees insight into their true level of skill if they were widely available. This instrument is designed to be unambiguous in the error descriptions to improve objectivity, but as a result errors must be directly observed on the arthroscopy screen:¹²⁸ off-camera errors or lapses in the assessor's attention reduce the detection accuracy. The one example identified is comprehensive (77 items) and would need significant attention to complete accurately. Assessors required 8 hours of training to use the tool, and two assessors review each procedure in duplicate. This is clearly unfeasible for regular use. Because the tool measures (hopefully) increasingly infrequent events as training progresses, it's effectiveness at higher skill levels must be questioned on the grounds of simple probability* and the likely decline of vigilant attention when the detection of errors is low.²⁶⁴ As only one error list was identified for use in a simulated setting, it is unclear how effectively the concept translates to the operating theatre or other procedures.

Table 2.12: Evidence demonstrated for application of error lists as quantitative performance metrics.

Paper	Error list	Application	KP	Summary of findings
Angelo ¹⁹⁸ 2015 Arthroscopy	Copernicus (77 errors, with 20 sentinel errors)	Learning	NT	Time taken, errors and sentinel errors made showed statistically significant differences with operative experience, but no difference in the surgical steps
Angelo ¹²⁹ 2015 Arthroscopy	Copernicus (77 errors, with 20 sentinel errors)	Learning	2	Surgical steps, errors and sentinel errors made statistically significantly improved with proficiency-based training, with no differences in time taken
Angelo ¹⁹⁹ 2015 Arthroscopy	Copernicus (77 errors, with 20 sentinel errors)	Learning	NT	Time taken and errors made showed statistically significant differences with operative experience, but no difference in the surgical steps or 'sentinel' errors

Continued on next page

*A participant with a 10% sentinel error rate would go undetected when assessed about one in eight times: $(1 - 0.1)^{20} \approx 12\%$

Table 2.12: Evidence demonstrated for application of error lists as quantitative performance metrics.

Paper	Error list	Application	KP	Summary of findings
Angelo ¹²⁸ 2015 Arthroscopy	Copernicus (77 errors, with 20 sentinel errors)	Reported	NT	Face and content validation of error checklist metrics

KP: Kirkpatrick level demonstrated in study overall. NT: Non-training intervention

2.4.8 Behavioural features

This groups quantitative metrics where performance is assessed by recording events associated with inexperience. Unlike error lists, where events are likely to directly cause iatrogenic harm or treatment failure,¹²⁸ these metrics count the frequency of events which aren't necessarily harmful in isolation.

Fifteen studies reported the frequency of behavioural features as a quantitative performance measure. The commonest metric was the frequency with which the arthroscopic instruments collided with tissues or other instruments in a VR setting, reported in ten studies. Four of those studies also measured the force of the collisions to further characterise 'injurious' tissue collisions^{197,233,238,239} (see Force Parameters below). Using automated measures of camera-handling behaviour, Fucentese²¹³ and Gandhi²¹⁴ did not show learning, but using external observers Alvand¹⁹⁴ showed learning in a simulated setting where more experienced participants lost instruments on-screen less often, and took less time to find lost instruments again. Alvand¹⁹⁴ and Cowan⁵¹ both used external observers to show more experienced participants looked down at their hands less frequently than novices during a procedure. Table 2.13 summarises the evidence demonstrated for the application of behavioural features as quantitative performance metrics and overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 202).

Recording behavioural features attempts to quantify what might be used subjectively to justify grading inexperience-experience in a global rating scale. Their objectivity is enhanced by being unambiguously described (like surgical errors¹²⁸) reducing the requirement for interpretation and the potential for bias from observers, or by automating the process in the case of VR simulators. When considering

using behavioural features to monitor in the operating theatre and to link to PROMs, it is important to note that for the behavioural features where performance has been shown to improve with experience (e.g. look downs, tissue/instrument collisions, triangulation time and dropped items) either require dedicated observers, or technology for collision detection that doesn't exist in the real world. The requirement for observers could be circumvented by improvements to computer vision technology which may in the future automate, for example, the frequency of instrument loss and triangulation time, but this is has not been assessed by any of the retrieved studies.

Table 2.13: Evidence demonstrated for use of behavioural features as quantitative performance metrics.

Paper	Behaviour [method]	Application	KP	Summary of findings
Alvand ¹⁹⁴ 2012 JBJS Am	Look downs', instrument loss, triangulation time [Observer]	Learning	NT	Good correlations ($r = 0.69$ to 0.88) of performance by simple visual parameters with GRS & motion analysis metrics
Andersen ¹⁹⁷ 2011 Acta Orthop	Tissue collisions [VR - insight-MIST]	Reported	2	Time taken, path length, and collision depths statistically significantly improve with training.
Bliss ²⁰¹ 2005 SHTI	Tissue collisions [VR - Pro-cedicus]	Reported	2	Landmark identification, time taken and tissue collisions showed statistically significant improvement over training period. 'Minimal' degradation/skills decay over 4 weeks, but no statistics reported
Braman ²⁰² 2015 Arthroscopy	Accidental contacts, dropped pegs [Audio & visual alarms]	Learning	NT	Time taken, accidental contacts and dropped pegs show statistically significant improvement with operative experience on some simulated tasks
Chang ²⁰⁷ 2016 CORR	"Care for soft tissues" [Undocumented]	Reported	NT	Time taken, GRS and global score shows statistically significant improvement with operative experience. 'Care for soft tissues' results not reported.
Cowan ⁵¹ 2016 Orthopedics	Look downs, chondral injuries [Observer & VR - ArthroSim]	Learning	NT	All trainees susceptible to distraction, but senior trainees still require less time, cause fewer injuries, and look at hand less often
Fucentese ²¹³ 2015 KSSTA	Camera horizon alignment [VR - ArthroS]	Reported	NT	Time taken and path length show statistically significant improvement with operative experience, but no difference in camera horizontal alignment
Gandhi ²¹⁴ 2015 Arthroscopy	Movements outside of working area [Computer]	Reported	NT	Correlations between performance on internet-based tool and VR simulator
Gomoll ²¹⁷ 2007 AJSM	Tissue collisions [VR - Pro-cedicus]	Learning	NT	Statistically significant improvements in time taken, collisions, path length and probe velocity with operative experience

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.13: Evidence demonstrated for use of behavioural features as quantitative performance metrics

Paper	Behaviour [method]	Application	KP	Summary of findings
Gomoll ²¹⁸ 2008 AJSM	Tissue collisions [VR - Pro- cedicus]	Learning	2	Time take, collisions, velocity and path length statistically significantly improved after 3 years of clinical training, equivalent performance to matched historical group.
McCarthy ²³⁰ 2006 CORR	Instrument collisions [VR - SKATS]	Reported	NT	Time taken, arthroscope path length and number of loose bodies identified showed statistically significant improvement with operative experience.
Pedowitz ²³³ 2002 Arthroscopy	Tissue collisions [VR - Pro- cedicus]	Reported	NT	Time taken and path ratio improve with operative experience, but no change in force parameters
Smith ²³⁸ 1999 SHTI	Tissue collisions [VR - Pro- cedicus]	Learning	NT	Trained orthopaedic surgeons take less time and make fewer collisions than medical students and other surgeons across a range of exercises
Srivastava ²³⁹ 2004 JSES	Tissue collisions [VR - Pro- cedicus]	Learning	NT	Time taken and surgical navigation score statistically significantly improve with operative experience.
Strom ¹⁸⁷ 2004 Surg Endosc	Tissue collisions [VR - Pro- cedicus]	Reported	2	No effect of visuospatial training on time taken, path length, collisions or overall score during virtual arthroscopy noted

KP: Kirkpatrick level demonstrated in study overall; NT: Non-training intervention

2.4.9 Force parameters

This groups quantitative metrics where performance is assessed by any measure of force (N) on tissues or instruments.

Eleven studies measured force as a quantitative measure of technical performance, with four demonstrating differences with training or experience in a simulated setting.^{211,232,242,265} Chami et al⁷¹ used an instrument-mounted force sensor, displaying the force profile on-screen during an intra-operative diagnostic knee arthroscopy, showed different patterns in the “noise/task features” of experts, who have fewer noise features for every task feature than trainees, but this was not correlated to patient outcomes. Beyond the general force profile during a procedure, five studies^{197,232,233,238,239} attempted to define a force threshold for injurious tissues collisions, but only one²³² demonstrated a performance difference with training or experience. Three studies^{211,212,242} attached benchtop simulators to force transducers to illustrate how force and torque applied to the limb during knee arthroscopy

decrease with experience, and two studies^{216,265} force tested arthroscopically tied knots, though only Pedowitz et al²⁶⁵ showed a learning effect, where a short training intervention with feedback improved the strength of the knot. Table 2.14 summarises the evidence demonstrated for the application of force as a quantitative performance metric and overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 203).

The controlled application of force is important in surgery, particularly orthopaedics. The ability to measure forces applied to intra-articular tissues and visualise this intra-operatively on the arthroscopy screen, as shown by Chami et al, could be both a guide and performance metric of iatrogenic harm if combined with appropriate thresholds for injurious tissue contacts.²³² Other direct force measures, such as attaching the patient's limb to force transducers, are likely to be cumbersome and less informative than intra-articular force measurement despite indicating they are useable to distinguish participants based on experience. As well as direct measurement by force transducers, force was attempted to be inferred by Kirby et al by measuring the maximum acceleration which was anticipated to correlate to more forceful motions,²²³ but the effect of gravity dominates the accelerometer readings meaning only movements downward trigger a peak reading. With the appropriate medical devices regulations in place, and if the sensitive equipment can be made to withstand sterilisation (or be made cheap enough for single use), intra-articular force measurement is potentially a useful quantitative measure of arthroscopic performance, but these few studies show it is a less well validated field than motion parameters.

Table 2.14: Evidence demonstrated for application of force parameters as quantitative performance metrics

Paper	Forces [method]	Application	KP	Summary of findings
Andersen ¹⁹⁷ 2011 Acta Orthop	Depth of tissue collisions [VR - insightMIST]	Reported	2	Time taken, path length, and collision depths statistically significantly improve with training.
Chami ⁷¹ 2008 CORR	Instrument force and torque [Instrument-mounted transducer]	Behaviour	NT	Correlation between operative experience with time taken and force parameters

KP: Kirkpatrick level demonstrated in study overall; NT: Non-training intervention. Continued on next page

Table 2.14: Evidence demonstrated for application of force parameters as quantitative performance metrics

Paper	Forces [method]	Application	KP	Summary of findings
Escoto ²¹¹ 2012 SHTI	Joint force and torque [Simulator mounted on force plate]	Learning	NT	Time taken and average force applied improved with operative experience
Escoto ²¹² 2013 IEEE	Joint and instrument force and torque [Simulator mounted on force plate and instrument sensors]	Reported	2	Apparent improvement in time taken and path length performance after practice, but no statistical confirmation. Device able to measure forces, but not results not reported.
Gilmer ²¹⁶ 2015 Arthroscopy	Knot strength [load testing to failure]	Reported	2	Knot-tying speed and confidence improve with training, but no change in knot strength
Nicandri ²³² 2015 OJSM	Probe pressure [VR - TolTech]	Learning	2	Time taken, arthroscope and probe path length, smoothness and probe pressure improved with practice on the simulator. No significant change in completeness of procedure.
Pedowitz ²³³ 2002 Arthroscopy	Injurious tissue collisions [VR - Proceidicus]	Reported	NT	Time taken and path ratio improve with operative experience, but no change in force parameters
Pedowitz ²⁶⁵ 2015 Arthroscopy	Knot strength [static load testing]	Learning	2	Proficiency training reduces knot failure rates
Smith ²³⁸ 1999 SHTI	Injurious tissue collisions [VR - Proceidicus]	Reported	NT	Trained orthopaedic surgeons take less time and make fewer collisions than medical students and other surgeons across a range of exercises
Srivastava ²³⁹ 2004 JSES	Injurious tissue collisions [VR - Proceidicus]	Reported	NT	Time taken and surgical navigation score statistically significantly improve with operative experience.
Tashiro ²⁴² 2009 CORR	Joint force and torque [Simulator mounted on force plate]	Learning	NT	All metrics show statistically significant improvement with operative experience

KP: Kirkpatrick level demonstrated in study overall; NT: Non-training intervention

2.4.10 Single-item scales

This groups quantitative metrics where performance is assessed by response to a single-item Likert or visual analogue scale, by either the surgeon or an observer.

Six studies used a single-item scale to assess performance: three report the safety^{185,186} or harm²³⁶ caused, two report overall skill,^{252,266} and one study reports adequacy of meniscal resection.²⁴² The Waterman and Dunn studies appear to be drawn from a similar cohort at differing recruitment stages.^{185,186} Both report the ASSET Safety domain as an individual metric isolated from the ASSET global rating scale and checklist. Waterman et al report an intra-operative improvement in performance during a diagnostic shoulder arthroscopy after an hour of VR training

using the ASSET Safety score, but no improvement was noted by Dunn et al. Two studies^{216,232} reporting participants assessment of their confidence on a single-item scale were not included because the studies ask participants about confidence in their ability to perform the procedure in general, rather than their confidence during a given attempt. Table 2.15 summarises the evidence demonstrated for the application of single-item scales as a quantitative performance metric and overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 203).

Single-item scales share similarities with global rating scales and calculated scores (see below, section 2.4.12), but are less common according to this search. These three methods all report performance as a single value, but while global rating scales use summed subjective rankings across a broader set of defined domains and calculated scores apply a weighted contribution from various data sources to generate a single summary measure, the single-item scales are a subjective summary without the intervening steps. Without the prompts from a global rating scale to consider various skill domains, a single-item assessment of skill are more prone to bias. This is particularly important in participant self-rating²⁶⁶ where it is unclear whether junior surgeons can take an objective view of their own performance^{175,215} and an unawareness of the required skill domains could lead to “ignorant incompetence”.^{176,267}

It is difficult to assess the value of single-item scales of arthroscopic performance because none of the studies found validated the construct of the single-item scale used, or correlated it with other validated metrics. Global rating scales gain some of their validity from the internal consistency between the domains,²⁵⁵ and using a domain in isolation such as the ASSET Safety score^{185,186} is dubious. Moreover, the Safety domain had the poorest inter-rater reliability compared to other domains in the ASSET global rating scale in its initial validation.^{224,249} Waterman et al report intra-operative improvements after VR training in the ASSET Safety domain, but not the overall ASSET global rating, and the usefulness of those findings should be viewed in this context.

Table 2.15: Evidence demonstrated for application of single-item scales as quantitative performance metrics.

Paper	Scale description [rater]	Application	KP	Summary of findings
Dunn ¹⁸⁶ 2015 J Surg Educ	ASSET Safety domain; anchored, 1-5 [Observer]	Reported	3	Time taken and GRS performance improves in both control and experimental groups, but no differences between the two groups. Task specific checklist and single item scale performance show no change.
Rebolledo ²³⁶ 2015 AJSM	Injury Grading Index; anchored, 1-10 [Observer]	Learning	2	Simulator training improved time and IGI performance in shoulder model, but not knee model
Slade Shantz ²⁵² 2013 Arthroscopy	Resident skill; VAS, “Has no skill” to “Extremely skilled” [Observer]	Reported	NT	GRS and single-item VAS show statistically significant improvement with operative experience
Tashiro ²⁴² 2009 CORR	Adequacy of resection of meniscal tear; anchored, 1-4 [Observer]	Learning	NT	All metrics show statistically significant improvement with operative experience
Unalan ²⁶⁶ 2010 KSSTA	Self skill rating; 1-10 [Participant]	Learning	2	Improved self-confidence over course
Waterman ¹⁸⁵ 2016 Orthopedics	ASSET Safety domain; anchored, 1-5 [Observer]	Behaviour	3	In theatre, only improvement in Safety domain of ASSET after training (which has low reliability in original study). In simulator, trained group better time and path length.

KP: Kirkpatrick level demonstrated in study overall; NT: Non-training intervention

2.4.11 Virtual reality proprietary metrics

This groups quantitative metrics where performance is assessed in a manner unique to the virtual-reality simulator, and were judged not to be easily independently calculable on another simulator or in theatre. This excludes summary scores, which are discussed below.

Six studies reported six different metrics specific to the VR simulator in use which have not been validated elsewhere. Cannon et al¹⁴⁵ assessed the flexion, valgus and varus angles of the knee during the simulation. McCarthy,²³⁰ Nicandri²³² and Rose²³⁷ reported variants of the position of anatomical features or instruments on the screen. Strom¹⁸⁷ reported the path-length of the instruments relative to the optimum straight-line path-length between points, and Srivastava²³⁹ reported the number of sequentially randomly generated objects within the joint that were identified in a 5-minute period by the participant. Table 2.16 summarises the evidence demonstrated for the application of VR proprietary metrics as quantitative

performance metrics and overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 204).

In the real-world setting, calculating a patient-specific optimum path-length and random generation of intra-articular bodies are clearly not possible. The live measurement of joint angles could be managed by measuring relative position of the thigh and calf, although this introduces electromagnetic fields and sensors into the operating room and compromises sterility, while the automated reporting of on-screen events would require significant improvements in computer vision technology. Though these VR proprietary metrics could in due course be applied to the operating theatre and be used to investigate intra-operative behaviour and how they affect patient outcomes, they are currently impractical and expensive, and other better validated measures of performance are available.

Table 2.16: Evidence demonstrated for application of VR proprietary metrics as quantitative performance metrics.

Paper	Proprietary metric	Application	KP	Summary of findings
Cannon ²⁰⁵ 2014 JBJS Am	Flexion, valgus, and varus angle [ArthroSim]	Reported	NT	Time taken shows statistically significant improvement with operative experience, but no differences in summary score.
McCarthy ²³⁰ 2006 CORR	Probe observation [SKATS]	Reported	NT	Time taken, arthroscope path length and number of loose bodies identified showed statistically significant improvement with operative experience.
Nicandri ²³² 2015 OJSM	Completeness [TolTech]	Reported	2	Time taken, arthroscope and probe path length, smoothness and probe pressure improved with practice on the simulator. No significant change in completeness of procedure.
Rose ²³⁷ 2015 Arthroscopy	Tracking moving target [ArthroVision]	Learning	NT	Inconsistent support for better performance with operative experience across three tasks with multiple metrics.
Srivastava ²³⁹ 2004 JSES	Surgical navigation [Procedicus]	Learning	NT	Time taken and surgical navigation score statistically significantly improve with operative experience.
Strom ¹⁸⁷ 2004 Surg Endosc	Optimum path length [Procedicus]	Reported	2	No effect of visuospatial training on time taken, path length, collisions or overall score during virtual arthroscopy noted

KP: Kirkpatrick level demonstrated in study overall; NT: Non-training intervention

2.4.12 Calculated scores

This groups quantitative metrics where performance is assessed using a single summary score, calculated from weighting the contribution of other recorded metrics and presenting the summary performance as a grade or an integer relative to a maximum score. This includes summary ‘global scores’ commonly reported by VR simulators.

Scores were reported by 12 studies, with 6 demonstrating an improvement in the score with training or experience. The manually calculated scores^{91,210,223,235,246,255,268} all had transparent inputs and transparent calibration of the contribution of each input to generate the score. Kirby et al,²²³ Jacobsen et al⁹¹ and Rahm et al²³⁵ presented evidence for the rationale behind their method of calibrating the score, Kirby using machine learning and Jacobsen and Rahm using statistical normalisation of other metrics, but both of these methods require all data from all participants to be acquired before calculating the scores for each participant, so cannot be used in an *ad hoc* fashion for individual training events. The other manually calculated scores all used a variation of a ‘time score’ (completing the procedure relative to an optimum time), with a deducted ‘penalty score’ for any predefined errors which occurred, but did not provide evidence for how they calibrated their scores. Automated scores were all calculated by the VR simulator being used in the study (ArthoSim,²⁰⁵ ArthroVR^{207,215,269} or Proceidius¹⁸⁷), but only Strom et al¹⁸⁷ report the inputs contributing to the score, and none of the studies report the method or evidence behind calibrating the score. This is understandable, as this information makes the scores potentially ‘gameable’, and the intellectual property behind a validated ‘summary score’ which predicted intra-operative performance would be lucrative. Table 2.17 summarises the evidence demonstrated for the application of calculated scores as quantitative performance metric and overall Kirkpatrick level in the identified studies, with full study details available in Appendix B.2 (pg. 204).

Scores should be usable by any level of surgeon, as they attempt to make interpretation of technical performance easier by summarising multiple metrics which may be difficult to interpret all together. That poses two reciprocal problems.

Where opaque calibration of the score masks the contribution of the inputs, it makes training and deliberate practice more challenging as the parameters of success are hidden, whereas transparent calibration may reinforce bad habits by allowing a strategic decision to sacrifice performance in one domain and focus on improvements in a more favourably weighted domain to optimise their score. Transparency is important, especially if quantitative metrics are going to be used for making decisions about selection or progress through proficiency-based training: surgeons have a right to know how they are being judged, but that knowledge would mean a summary score is ‘gameable’. No summary score has been linked to improvements in intra-operative performance or patient outcomes, and some of the VR calculated scores will not be usable in the operating theatre. Given that the evidence behind the method of calibrating of each score is mostly unclear, it is difficult to justify summary scores being a fair or effective performance metric to persist with.

Table 2.17: Evidence demonstrated for application of calculated scores as quantitative performance metrics.

<i>Paper</i>	<i>Score inputs</i>	<i>Transparent</i>	<i>Calibration evidence</i>	<i>Application</i>	<i>KP</i>	<i>Summary of findings</i>
Bayona ²⁰⁰ 2008 J Robot Surg	Time, goals, errors, anatomy knowledge, steadiness, ambidexterity, roughness, dexterity, hand eye coordination, response to critical settings [Manual formula]	Yes	Expert opinion	Reported	2	Reports face and content validity of simulator. Score is described, with contribution of each metric weighted by expert opinions of it's importance, but not used to assess performance
Cannon ²⁰⁵ 2014 JBJS Am	Proper arthroscope position, probing technique, knee flexion angle, and correct valgus or varus force on the knee joint [Automated; VR - ArthroSim]	No	Unclear	Reported	NT	Time taken shows statistically significant improvement with operative experience, but no differences in summary score.

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention. Continued on next page

Table 2.17: Evidence demonstrated for application of calculated scores as quantitative performance metrics.

<i>Paper</i>	<i>Score inputs</i>	<i>Transparent</i>	<i>Calibration evidence</i>	<i>Application</i>	<i>KP</i>	<i>Summary of findings</i>
Chang ²⁰⁷ 2016 CORR	Unclear [Auto- mated; VR - ArthroVR]	No	Unclear	Learning	NT	Global score shows statistically significant improvement with operative experience. 'Care for soft tissues' results not reported.
Coughlin ²⁶⁸ 2015 JBJS Am	Maximum score, minus time taken and errors made [Manual formula]	Yes	Unclear	Learning	NT	Global score shows statistically significant improvement with operative experience
Elliot ²¹⁰ 2012 Arthroscopy	Time taken, chondral injuries, structures visualised [Manual formula]	Yes	Unclear	Learning	NT	Time taken, TSCL and ASAF-Knee scoring metric show statistically significant improvement with operative experience
Ferkel ²⁶⁹ 2013 Arthroscopy	Unclear [Auto- mated; VR - ArthroVR]	No	Unclear	Reported	NT	Poor correlation of VR global score to annual operative case load
Gandhi ²¹⁵ 2015 SECEC	Unclear [Auto- mated; VR - ArthroVR]	No	Unclear	Reported	NT	No statistically significant correlation between participants subjective scoring of their manual dexterity and hand-eye coordination with VR performance parameters
Jacobsen ⁹¹ 2015 JBJS Am	Time and motion [Manual formula]	Yes	Statistical model	Learning	NT	Reports construct validity with all metrics. Calculated score with the thresholds described would fail 1 of 13 qualified surgeons and pass 2 of 13 novices in high stakes assessment.

KP: Kirkpatrick level demonstrated in study overall; NT: Non-training intervention. Continued on next page

Table 2.17: Evidence demonstrated for application of calculated scores as quantitative performance metrics.

<i>Paper</i>	<i>Score inputs</i>	<i>Transparent</i>	<i>Calibration evidence</i>	<i>Application</i>	<i>KP</i>	<i>Summary of findings</i>
Kirby ²²³ 2015 JBJS Am	Motion parameters and time [Manual k-means clustering]	Yes	Machine learning	Learning	NT	Arthroscopic performance score from unsupervised machine learning technique more accurately defines the three experience groups when sensors worn at elbow, and correlates well with GRS. Elbow and wrist worn sensors recorded hand movements and smoothness show strong correlation.
Rahm ²³⁵ 2016 BMC Surg	Motion parameters and time [Manual formula]	Yes	Statistical model	Learning	2	Overall score improves with practice over first 2 hours of training, then plateaus
Strom ¹⁸⁷ 2004 Surg Endosc	Collisions, movements economy and time [Automated; VR - Procedicus]	No	Unclear	Reported	2	No effect of visuo-spatial training on time taken, path length, collisions or overall score during virtual arthroscopy noted
Wong ²⁴⁶ 2013 KSSTA	Time, tissue approximation, knot slippage/disruption [Manual formula]	Yes	Unclear	Reported	2	Self-calculated 'proficiency score' correlates well with observer-recorded GRS and TSCL

KP: Kirkpatrick level demonstrated in study overall; NT:Non-training intervention

2.4.13 Fluoroscopy

Fluoroscopy is an important tool for intra-articular access and surgical guidance in hip and wrist arthroscopy. Because fluoroscopy screening times have been associated with experience in a dynamic hip screw procedure,²⁷⁰ this was identified as a potential metric group to measure technical performance in arthroscopy. Two studies to include fluoroscopy data were identified at the initial search, but these were excluded from full-text review as they did not link fluoroscopy exposure to experience or performance, rather using it to quantify a radiation risk.^{271,272}

2.4.14 Correlation of quantitative measures of arthroscopic performance to intra-operative behaviour

Twenty studies were identified which had an in-theatre component, and 17 identified intra-operative behaviour differences using a quantitative measure of arthroscopic performance. Ten single-surgeon case-series^{126,179,196,206,208,209,222,225,245,273} used time taken to assess the ‘learning curve’ with increasing experience of a new arthroscopic procedure and one randomised controlled trial¹⁸⁶ used time taken to demonstrate intra-operative improvements with simulation training. Five studies showed intra-operative behavioural differences with a global rating scale, either assessing the impact of simulation training on intra-operative performance^{144,185,186} or during a construct validation study^{224,251} demonstrating the impact of previous training and experience on performance. Two randomised studies used task-specific checklists^{144,145} to show the effect of simulation training on intra-operative performance. One construct validation study used probing instrument force profiles⁷¹ and one randomised control trial¹⁸⁵ used ASSET Safety domain as a single item scale to show differences in intra-operative performance.

Three studies performed in theatre did not demonstrate performance differences with their chosen metrics: one recorded task specific checklist performance intra-operatively to validate the internal consistency and inter-rater agreement,²⁶³ one used a global rating scale to assess a new method of monitoring the learning curves in knee arthroscopy,²⁵⁸ and one study found no impact of telementoring software on procedure time.¹⁸⁴

Four studies were randomised controlled trials specifically assessing the impact of simulation training on the performance of surgical trainees during diagnostic knee or shoulder arthroscopy. Three studies showed that after a period of simulation training the intervention group scored better than the control group using a global rating scale,¹⁴⁴ task-specific checklist^{144,145} or single item scale,¹⁸⁵ demonstrating Kirkpatrick Level 3 response to simulation training (‘behavioural change’), while Dunn et al¹⁸⁶ demonstrated that both the intervention and control group improved

Table 2.18: Cochrane Risk Of Bias assessment for randomised trials of simulation training reporting statistically significant improvements in quantitative performance metrics.

	<i>Cannon 2014¹⁴⁵</i>	<i>Dunn 2015¹⁸⁶</i>	<i>Howells 2008¹⁴⁴</i>	<i>Waterman 2016¹⁸⁵</i>
Random sequence generation	Low	Unclear	Low	Unclear
Allocation concealment	Unclear	Unclear	Low	Unclear
Blinding of participants & personnel	Unclear	Unclear	Unclear	Unclear
Blinding of outcome assessment	Low	Low	Unclear	Low
Incomplete outcome data	Low	Low	Low	Low
Selective reporting	Low	High	Low	High
Other	Low	High	Low	High

time taken and GRS performance during the study, but there was no significant inter-group difference.

Risk of bias was assessed for these randomised controlled studies using the Cochrane Collaboration tool¹⁸³ (Table 2.18). None of the trials could blind the participants to their randomisation status. Assessors rated intra-operative performance by watching anonymised arthroscopic videos or live in theatre where they were blinded to the participants randomisation status but it is unclear if the supervisor could infer this from their peri-operative behaviour on the day. All studies have apparently complete data and do not record any participant withdrawals. Two of the four randomised trials were felt to be at high risk of bias due to selective reporting of outcomes and potential other bias. They present unusual subgroup analysis of the data which is not validated by the original description of the instrument they use to measure performance²²⁴ which may represent an attempt to dredge the data, and the studies appear to be drawn from the same cohort at different recruitment stages released as least publishable units. None of the trials were registered to check intended reporting and adherence to protocols. The four randomised trials were too divergent in methods and outcomes to consider meta-analysis.

2.4.15 Correlation of quantitative measures of arthroscopic performance to patient-reported outcome measures

Six studies recorded patient-reported outcomes and attempted to link these to arthroscopic performance.^{196,206,222,225,245,273} These studies were all single-surgeon case studies reporting performance as time taken to plot the learning curve as they attempted a new technique. None of these studies showed an effect of case-series number or surgical time on patient outcomes in experienced surgeons using a novel technique.

2.5 Discussion

2.5.1 Summary of findings

This study has identified 92 primary arthroscopic studies using quantitative methods to assess arthroscopic technical skill. The metrics in the studies were grouped into 10 themes based on common features to structure the analysis on the nature of the metrics themselves, and the implications of their use as measures of performance. Five classes of metric in 20 studies have been applied in the operating theatre (Time, Global rating scales, Task-specific check lists, Single item scales, and Force), but there was only limited evidence from four studies that linked improvement in-theatre performance to a training intervention. Improved performance was seen in trainees receiving simulation training in global rating scale, task-specific checklist scores, and single-item scales. Each of these three metrics is subjective, the task-specific checklists suffer from ceiling effects, and the single-item scale used is poorly validated. As such, the current evidence for arthroscopic simulation training leading to improved intra-operative performance (Kirkpatrick Level 3) may be overstated. Performance was not linked to patient outcomes in any study, and no meta-analysis was possible due to the heterogeneous settings and outcomes.

2.5.2 Strength and limitations

This systematic review was conducted along PRISMA guidelines to aid transparency and reproducibility. It retrieved more articles with a broader array of quantitative

performance metrics than earlier studies. The method used to classify the metrics identified had not been piloted before, but there was little ambiguity in classifying metrics. Other reviews have framed their results based on the setting, joint, or type of simulator used,^{274,275} while the use of metric classes focussed this assessment on the metrics themselves and allowed evidence for similar metrics to be considered together. Metrics were constructively critiqued to understand the limitations and biases associated with each of them, and understand how this might affect interpretation and the potential for perverse incentives if widely rolled out to training. Most importantly, this review distinguishes between *objective* and *quantitative* metrics, which have previously been treated as synonymous with significant implications for the way that certain data are analysed and interpreted.

This study had changes to the registered protocol, but these are reported and justified. The lack of a meta-analysis means that despite a large number of included participants this study did not benefit from a pooled estimate based on the metrics, because of the diversity of participant experience, joints studied, and methods of recording metrics. Some metrics such as time or certain global rating scales might be amenable to meta-analysis in the future if other sources of study variation are controlled. This study also used two unpiloted methods of grouping the studies: an extension of the concepts in the Kirkpatrick framework to interpret the same effects in non-training studies, and a novel method to theme the metrics. Neither of these encountered any ambiguity when assigning studies, and it has been stressed that describing the application of the metrics as ‘reported’, ‘learning’, ‘behaviour’, or ‘outcome’ is distinct from the Kirkpatrick levels. In retrospect, inter-rater agreement of this on a portion of the included studies would be beneficial.

2.5.3 In the context of the wider literature

This review deliberately did not consider quantitative measures of rare and delayed events. Rare events such as mortality or arthroscopic ‘conversion-to-open’ are insensitive performance metrics and are poorly suited to monitoring learning, and may even be dangerous as seen in the Bristol Heart case.¹²¹ Delayed metrics would

not be useful for actionable feedback during training. Other proxies such as PROMs or surgical process outcomes measure the interaction of the whole surgical team with the patient and are not representative of intra-operative activity.

The significant focus in the literature has been on construct validation: the ability of the metrics to distinguish participants based on technical skill. Little focus has been applied to the implications of using these metrics, any associated perverse incentives, and the impact of the “goals” implicitly set for trainees by monitoring a metric whether a competency threshold is defined or not.^{91,105} An expectation of how trainees will respond to a metric outside of a research setting when that metric may be used for decision-making about them is necessary before implementing it as a performance measure. For example, the thresholds set for indicative numbers of key surgical procedures a trainee has to have performed at completion of training have been shown to be both inadequate⁸⁴ and unachievable,^{83,123} and when implemented in trauma and orthopaedic training led to a perverse incentive for trainees to “unbundle” cases in their logbooks into smaller procedures which would artificially inflate the number of cases.¹²² When a metric becomes a target it ceases to be a good measure if surgeons are able to influence the metric to achieve a certain level of performance, or appear to achieve it, without achieving the behaviours the metric was intended to influence (“gaming the system”).

Quantitative assessment of performance in arthroscopy is behind the general surgical literature, who have investigated the feasibility of motion analysis in the operating theatre¹³¹ and linked global rating scales assessment to patient outcomes (complication rates) in bariatric surgery.²⁵ However, all branches of surgery are yet to use objective, quantitative performance metrics to demonstrate the transfer of simulation training into intra-operative performance.

2.5.4 Implications of this study

The purpose behind this study was to identify a quantitative method of assessing arthroscopic technical performance suitable for studying the transfer of simulation training into real improvements in intra-operative performance. That metric would

be validated to show it is a relevant and important measure of performance; is calibrated to show improvement with the acquisition of technical skill; can be delivered cost-effectively and reliably in during independent practice and assessment scenarios; provide useful, proximate feedback to trainees about how to improve; and ultimately is expected to correlate to patient outcomes.

Of the eleven metric classes identified, fluoroscopy was not assessed in any study and is typically restricted to hip and wrist arthroscopy practice, so should be discounted as a candidate metric. The calculated scores identified were either based on opaque metrics internal to a VR simulator which would preclude their use in theatre, or provided no evidence for their construction and weighting of their inputs. Similarly, other VR proprietary metrics and VR recorded behaviour features cannot be used in theatre and are discounted. The single-item scales either relied on an observer which is impractical for deliberate practice, or a self-rating which would only be a formalisation of a trainees current opinion of their ability. The identified scales had either not been validated or were shown to have poor inter-rater reliability. The items themselves are also inherently subjective, so this class would not be considered further.

Force parameters are relevant to care of intra-articular tissues and were feasibly recorded in theatre in one study, but are relatively underexplored compared to other metric classes and would require significant investment in sensor development and changes to intra-operative practice based on one study of only five participants.⁷¹ Behavioural features which depended on VR recording either require an observer or advanced computer vision technology which does not exist yet. One error list was identified: the concept is potentially very useful, but they are procedure-specific, observer-intense, and become increasingly insensitive as participants get more experienced.

Of the four remaining classes, Time, Motion parameters, Global rating scales, and Task-specific check lists, these are where the majority of the validation work has taken place (87/92 studies; 95%) and there is evidence to support these metrics being sensitive to improving technical skills. None of the studies identified assessed how

well these metrics were at providing feedback to permit deliberate practice, but their face and content validity assessment indicate they are relevant metrics. However, both global rating scales and task-specific checklists require an observer, highlighting their subjectivity and their ineffectiveness for independent deliberate practice. Time taken is a “crude and probably unacceptable”¹¹ measure of performance and may provide perverse incentive to trainees. It is worth noting that time taken was the only metric where any of the included studies attempted to correlate performance to PROMs, and no correlation was found in any of those 6 studies.

This leaves motion parameters. Efficiency of movement and dexterity are felt to be important measures of surgical technical skill.^{11,87,220,276} The concept has been well validated in the skills labs to correlate to acquisition of surgical skills, and distinguish between surgeons of different experience levels. In a recent study by Kirby et al,²²³ wireless elbow-worn inertial motion sensors are identified as a cost-effective solution which circumvents the previous technical difficulties preventing the use of motion sensors in theatre. Additionally, they can be used independently, allowing feedback for deliberate practice in the skills lab or operating lists without the need for an attentive observer.

Motion parameters represent the current best option for an objective, quantitative measure of arthroscopic performance. Further validation is needed to ensure the recently described novel elbow-mounted wireless inertial sensors are suitable to be considered for use in the operating theatre. Using motion analysis in theatre could objectively validate the transfer of simulation training to intra-operative performance, and be used to link performance to PROMS.

Simulation gives students permission to fail.

— Richard Satava

3

Validating objective quantitative assessment of technical skill using WASP

Contents

3.1	Introduction	75
3.1.1	The role of motion analysis in the assessment of surgical skill	75
3.1.2	Validating new methods of assessment	76
3.1.3	Aims	78
3.2	Study A - concurrent validity of wired and wireless motion analysis	78
3.2.1	Methods	78
3.2.2	Results	84
3.3	Study B - accuracy and reliability of wired and wire- less motion analysis	87
3.3.1	Methods	87
3.3.2	Results	89
3.4	Discussion	90
3.4.1	Implications of the studies	90
3.4.2	Limitations	95
3.4.3	Summary	96

3.1 Introduction

3.1.1 The role of motion analysis in the assessment of surgical skill

Manual dexterity is an accepted aspect of surgical technical skill.¹¹ It can be subjectively measured by tools such as global rating scales (GRS), task-specific check lists (TSCL), and sentinel error lists.^{87,128} These tools have been applied to many simulated and in-theatre tasks in arthroscopy and other specialties. However, they are time-consuming, rely on the presence of an engaged and qualified observer, and their subjectivity predisposes them to cognitive biases. Despite this, they are common in simulation research because they are cheap to use and can be applied to both simulated and in-theatre settings. They have been used to imply the transfer of simulated technical skills training to the operating theatre^{144,145} because these subjective, quantitative measures can be correlated to objective, quantitative measures of performance in the skills lab, such as motion analysis.^{223,231,277}

Motion analysis is used in the simulated setting to directly record surgical activity as an objective measure of manual dexterity. Electromagnetic,^{212,220,276} inertial,²²³ optical^{277,278} and mechanical²⁷⁹ motion analysis techniques have been applied to surgical simulation. They are more expensive and equipment-intense than GRSs and other subjective methods. The commonest technique uses electromagnetic devices: sensors attached to the surgeon or instruments calculate their relative position and orientation by changes in magnetic flux detected from three orthogonal coils within a fixed source, for example, the Polhemus Patriot system.²²⁰ This position data has been validated to provide information about the path length travelled by the surgeon's hands or instruments during the procedure, or can be converted to acceleration data with changes in acceleration being used to count the number of movements.^{220,276} Technical and sterility issues inhibit the use of wrist-worn electromagnetic sensors in theatre, so demonstrations of the impact of simulation training have relied on surrogate measures in theatre.

The 'wireless assessment of surgical performance' (WASP) is a system described by Kirby et al where wireless inertial sensors are placed near the surgeon's elbows



Figure 3.1: Wireless elbow-worn motion sensors being used while scrubbing before an arthroscopic procedure

as a reliable assessment of hand dexterity.²²³ They measure acceleration directly in 3 translational and 3 rotational axes by orthogonal accelerometers and gyroscopes, which is used to calculate counts of movements and other metrics. They are positioned at the elbow, outside the sterile field, and wirelessly stream data to a computer via Bluetooth (Figure 3.1, and Appendix A). The advantage of wireless elbow-worn sensors over wired wrist-worn sensors is that they can be used equivalently in the skills lab and in theatre to demonstrate the transfer of simulation training using the same objective metrics.

3.1.2 Validating new methods of assessment

Methods of assessment derived their value by demonstrating their validity: that they assess what they claim to be assessing.¹⁶⁰ Validity is not a single factor that a device, an assessment tool, or a simulation can ‘have’ independently, but is a property of the inferences we make supported by a spectrum of evidence with multiple facets.¹⁵⁹ Because of this, the validity of an assessment must be considered in the context of how it is being applied: ceiling and floor effects of GRSs and TSCLs show tools may display different degrees of validity depending on the context.^{262,280,281}

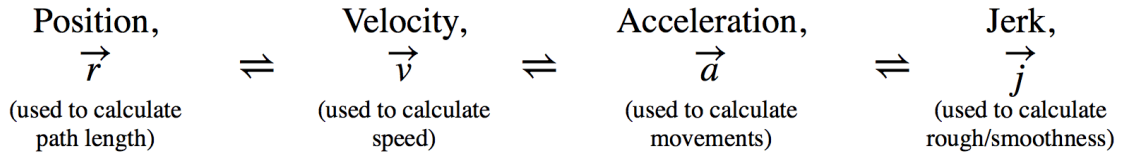


Figure 3.2: The relationship between the motion data values recorded by motion analysis devices. Electromagnetic devices record position, and inertial devices record acceleration. Lower order motion data, such as position, can be differentiated with respect to time to generate higher order derivatives, such as acceleration, to calculate metrics. Although mathematically possible, integrating higher order data into lower order data is sensitive to gravity and the accumulated errors in starting conditions making any calculated metrics very inaccurate e.g. integrating acceleration into position data gives inaccurate calculations for path length.²⁸²

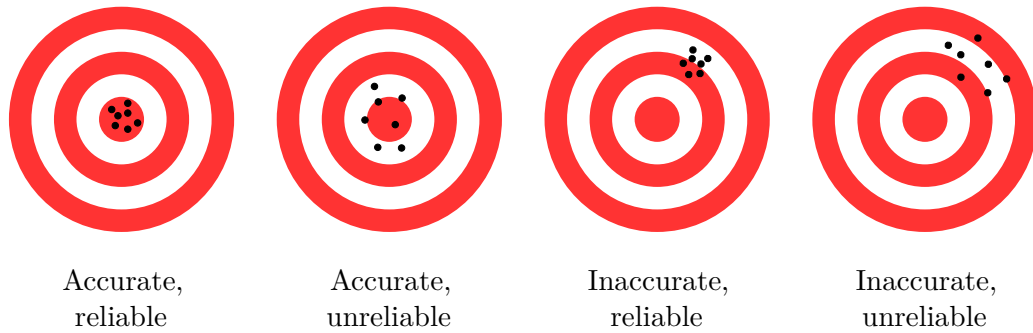


Figure 3.3: Accuracy and reliability of a metric are two independent features that affect results. Depending on the intended use different degrees of accuracy (bias) and reliability (precision) may be acceptable.

Electromagnetic and inertial devices can both be used as motion analysis devices. They calculate similar metrics, but using different methods and variables so they cannot be assumed to be interchangeable (Figure 3.2). Concurrent validity is the degree that two assessment methods like these measure the same skill domain (in this case, manual dexterity). Metrics must also be accurate, by being close to the true result, and reliable, by being repeatable with minimal variation (Figure 3.3). Reliability and accuracy both contribute to the degree confidence that a metric represents real life, which affects acceptance by users.

Demonstrating validity of the WASP motion analysis system is essential to justifying its use in the simulated and surgical setting, and to objectively demonstrate the transfer of simulated skills to the operating theatre.

3.1.3 Aims

The aim of this chapter is to assess the appropriateness of using a wireless system in place of the commonly used wired system to record hand movements. This was achieved by two linked experiments conducted using both systems simultaneously.

Study A to assess the concurrent validity of the metrics recorded by wireless and wired motion analysis devices during typical simulated arthroscopic use

Study B to assess the reliability and accuracy of the metrics recorded by wireless and wired motion analysis devices during stereotyped arthroscopic use

3.2 Study A - concurrent validity of wired and wireless motion analysis

3.2.1 Methods

Ethical approval

The study was conducted under ethical approval from University of Oxford Central University Research Ethics Committee (CUREC) (MSD-IDREC-C1-2014-152; Appendix A.4, pg. 191).

Participants

Participants were invited from surgeons at the Nuffield Orthopaedic Centre and students at the University of Oxford Medical School. Participants were selected to ensure a range of arthroscopic experience and ability (Table 3.1). No participants had previous experience of this simulated task. Participants signed informed consent forms before beginning the study. Outcomes were not shared with their supervisors, and had no impact on their training. The study was performed at the Oxford Orthopaedic Simulation and Education Centre (OOSSEC), at the Nuffield Orthopaedic Centre.

Table 3.1: Study A participant characteristics

Participant	Sex	National training number	Postgraduate year	Shoulder arthroscopies performed
A	Male	No	0	0
B	Male	No	2	0
C	Female	Yes	6	<10
D	Male	No	8	20
E	Female	Yes	10	39

Study setting

Simulated arthroscopy was performed in the skills laboratory at OOSEC using a standard 30° arthroscope, Smith & Nephew 460H camera, light source, and image management system, with images displayed on a 15-inch Iiyama TFT monitor. The equipment and simulation station arrangement is described in Appendix A.

Arthroscopic task

This study used a validated simulated diagnostic shoulder arthroscopy task to represent typical simulated arthroscopic activity.²⁸³ Participants viewed an orientation PowerPoint presentation describing the steps in the task before beginning the study. The task began with participants at rest with their hands in a standardised position on the shoulder model (Alex III Shoulder Professor, #1410-4-13; Sawbones, Malmö, Sweden) and the instruments on the table to each side of the model (arthroscope to the right; probe to the left). Once motion recording was started on both wired and wireless devices, the participant picked up the instruments and introduced the arthroscope into a posterior portal and the probe into an anterior portal of the simulated left shoulder (Figure 3.4). The goal was to visualise and probe nine clinically relevant landmarks (Table 3.2, and Figure 3.5). Once they had probed each landmark, they removed the arthroscope and probe, put them back on the table, and placed their hands on the shoulder model, then the motion recording was stopped. This was not a test of a recall or anatomy: participants could refer to a reference model which outlines the numbered landmarks at any point (Figure 3.5c & 3.5d). Beyond the procedural steps outlined in the orientation, participants were

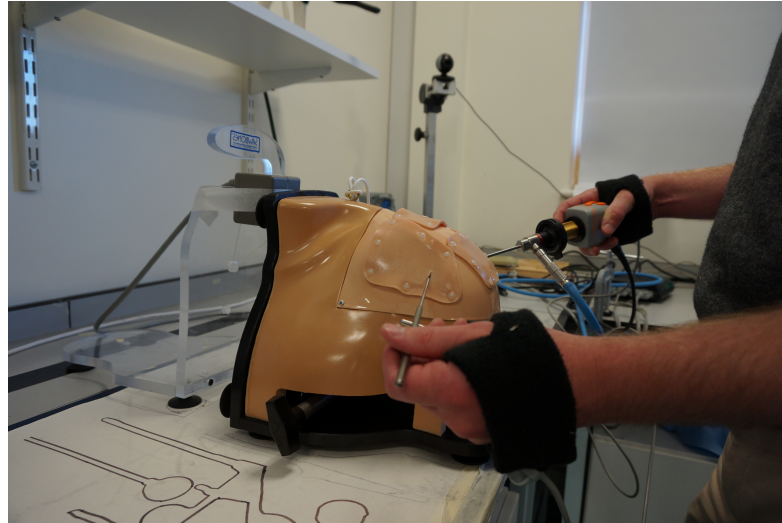


Figure 3.4: External view of the arthroscopic shoulder simulator in use. Note the wired wrist-worn motion sensors, and their electromagnetic source.

Table 3.2: Location and sequence of numbered landmarks to probe during simulated diagnostic shoulder arthroscopy

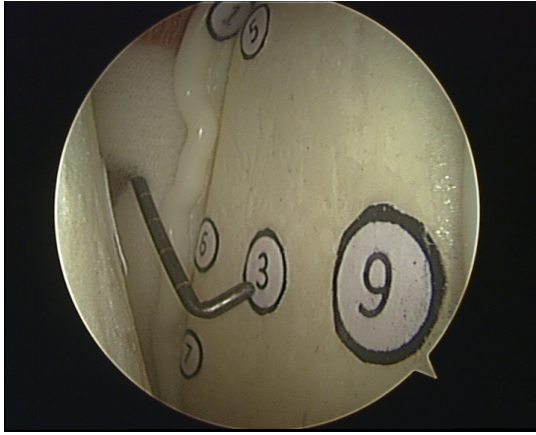
Step sequence	Landmark
1	Superior labrum
2	Centre of humeral head
3	Centre of glenoid
4	Superior humeral head, close to supraspinatus footprint
5	Superior glenoid - 12 o'clock
6	Anterior glenoid - 3 o'clock
7	Anterior glenoid - 4:30 o'clock
8	Anterior glenoid - 6 o'clock
9	Posterior glenoid - 9 o'clock

asked to take the same amount of care as they would with a real patient, and to complete the task to the best of their ability in an accurate, efficient and timely manner. Each participant completed the task five times while wearing both the wrist-worn wired motion sensors and elbow-worn wireless sensors.

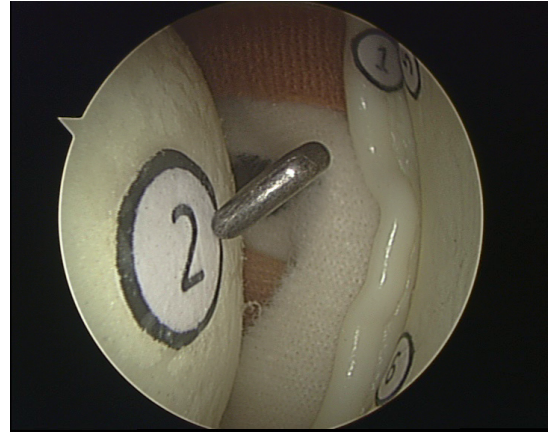
Motion analysis devices and metrics

The wired (Patriot) and wireless (WASP) motion analysis systems were used as discussed in the Introduction to this chapter, with further details in Appendix A. Figure 3.6 demonstrates the relative equipment burden of each system.

The Patriot system (Polhemus; Colchester, VT, USA) is an electromagnetic posi-



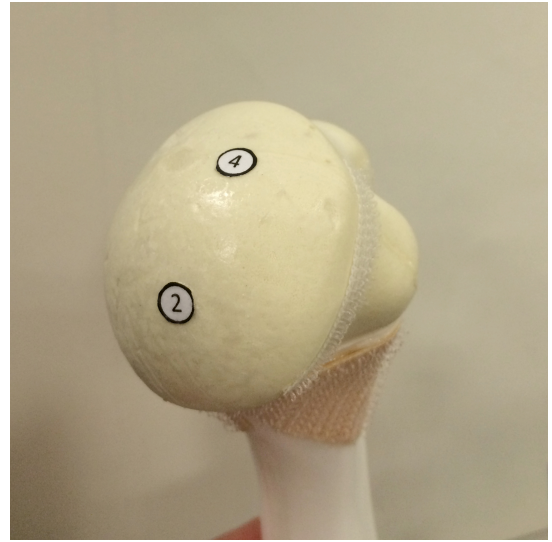
(a) Glenoid, arthroscopic view



(b) Humerus, arthroscopic view



(c) Glenoid, exposed view



(d) Humerus, exposed view

Figure 3.5: Arthroscopic and exposed views of the shoulder simulator landmarks.

tional motional analysis device, which has been well-described in the literature.^{110,127,220,283} It uses two wrist-worn electromagnetic sensors connected to a computer via the serial port, and an electro-magnetic field generator source located just behind the simulator in these tasks. Position data is acquired in 3 axes at a sampling frequency of 30Hz, and the static sensitivity is 10^{-3} m. Orientation data are not available on this system. The system was operated by a desktop computer. Data are acquired through the wired serial port via the command line interface, and manually analysed in MatLab on the desktop using custom scripts.²²⁰ These scripts report the path

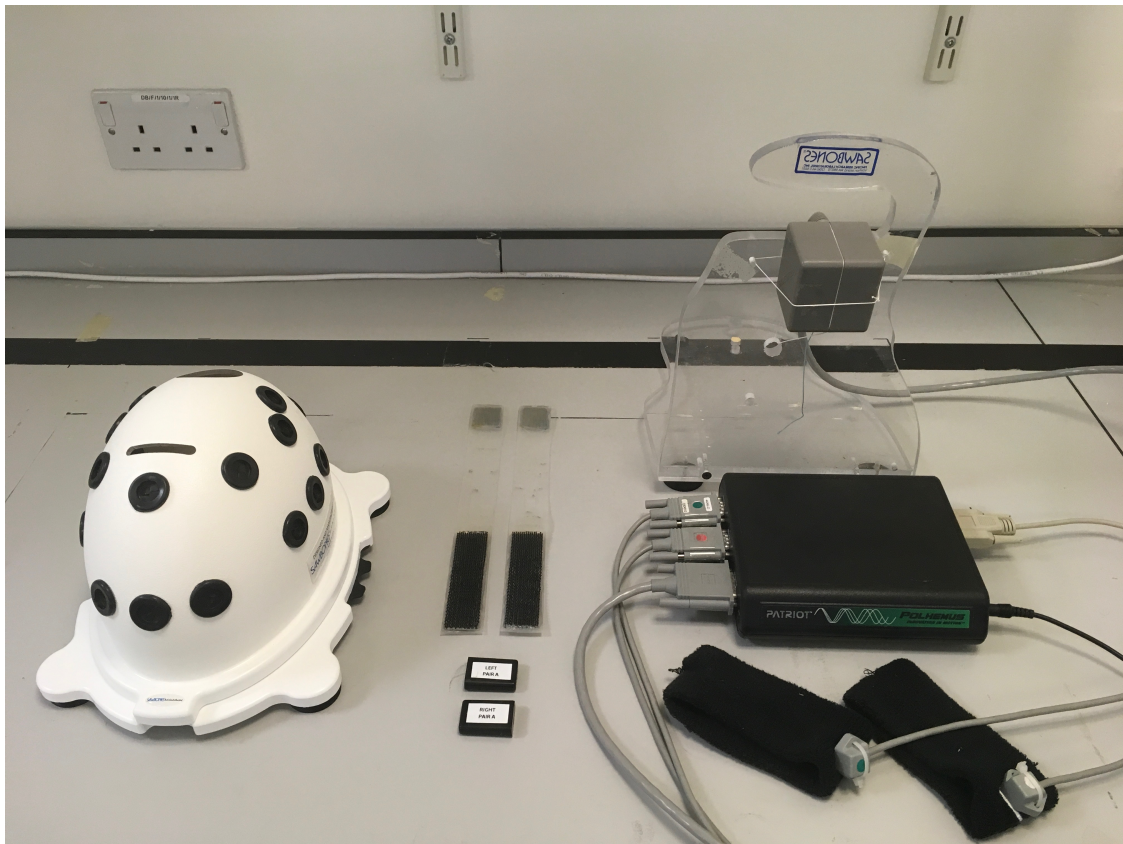


Figure 3.6: A comparison of the relative size of the motion analysis systems. In the center foreground, there are two wireless elbow-worn WASP sensors, with straps behind them which attach them to participants. To the right, in the foreground are the wired wrist-worn Patriot sensors attached to black wrist straps, and behind them the wired serial receiver and electromagnetic source. To the left, the FAST workstation is presented for comparison. Both systems also require a computer to receive and analyse their data.

length, number of hand movements, and time taken.

The wireless assessment of surgical performance system (WASP; McLaren Applied Technologies, Woking, UK) as described by Kirby et al is an inertial motional analysis device.²²³ Off-the-shelf WAX-9 inertial measurement units (Axivity, York, UK) are worn at the elbow just distal to each lateral epicondyle, over the extensor muscles. They record translational and rotational acceleration in three orthogonal axes directly with a sampling frequency of 50Hz, acceleration sensitivity of 0.1mg, and gyroscopic sensitivity of ± 2000 dps. Additional temperature, barometric pressure and 3-axis magnetometer sensors in the device were not used. The data are streamed wirelessly over Bluetooth via a custom client (WASP client; McLaren Applied

Technologies, Woking, UK) running on a laptop, then passed to an off-site web server and where they are analysed in MatLab using custom scripts. The results are returned to the user via the client. The scripts report two metrics in common with the Patriot wired system: number of movements, and time taken. Other available metrics can be found in the Appendix A.

Position, as measured by the Patriot system, measures the net effect of translational and rotational acceleration (surge, heave, and sway; and pitch, roll, and yaw), which are measured directly by the WAX-9 sensors. When calculating acceleration from positional information, the distinction between translational and rotational acceleration is lost. The WASP system had the option to use translational acceleration data (from accelerometers) with or without rotational acceleration data (from gyroscopes) when calculating the count of hand movements. Both methods were included for comparison in this study, and referred to as the wireless (ACC) system when discussing the wireless system using the accelerometer data alone, or wireless (ACC+GYRO) system when discussing the wireless system using the accelerometer and gyroscope data together.

Statistical analysis

Concurrent validity of the wired and wireless systems was assessed by the correlation of the count of hand movements required to complete the task reported by each device. The two devices record separately during the task, so the correlation for time taken would be used to ensure recordings of task attempts were similar. Spearman's rank correlation coefficient was used to assess correlation between the counts reported by each system as the data were not expected to be normally distributed. The wired system results were compared to the wireless (ACC) and wireless (ACC+GYRO) system results. Bland-Altman (mean-difference) plots were constructed for wired and both wireless systems reported count of hand movements to assess the agreement between the methods. Descriptive statistics (mean, SD, and range) for the wired and both wireless system methods were calculated. All statistical analyses were performed in R v3.3.2.²⁸⁴

3.2.2 Results

Five participants completed five attempts each at a simulated diagnostic shoulder arthroscopy. One data point was lost as the wired system results were found to be corrupted from a single attempt by participant B. It was removed, and 24 attempts were included in final analysis. The reported time taken by the wired and wireless devices showed a strong positive correlation (Spearman's $r_s = 0.99$, $p < 0.001$), indicating the recording periods were well-matched, and thus likely to include all movements and represent the whole task.

Comparison using only accelerometer data from wireless devices

The counts of hand movements calculated using the wireless (ACC) system was positively correlated with the wired system (Spearman's $r_s = 0.78$, $p < 0.001$; see Figure 3.7). The wireless (ACC) system, the wireless system tended to count fewer movements than to the wired system, with mean difference in the counts reported indicating a bias towards the wired devices (mean = 59.7), and a relatively broad distribution of the differences of counts reported (S.D. = 30.2) given the range of the means of the counts of movements in this study (range: 27, 108.5; see Figure 3.8).

Comparison using accelerometer and gyroscope data from wireless devices

The count of hand movements calculated using wireless (ACC+GYRO) system was more strongly positively correlated with the wired system (Spearman's $r_s = 0.91$, $p < 0.001$; Figure 3.9) than when using accelerometer data alone. The wireless system tended to count more movements when gyroscope data was included, with the mean difference in the counts reported indicating a bias away from the wired devices (mean = -28.3), and a narrower distribution in the differences of counts reported (S.D. = 13.1) given the range of the means of the counts of movements (range: 56.5, 209.5; see Figure 3.10).

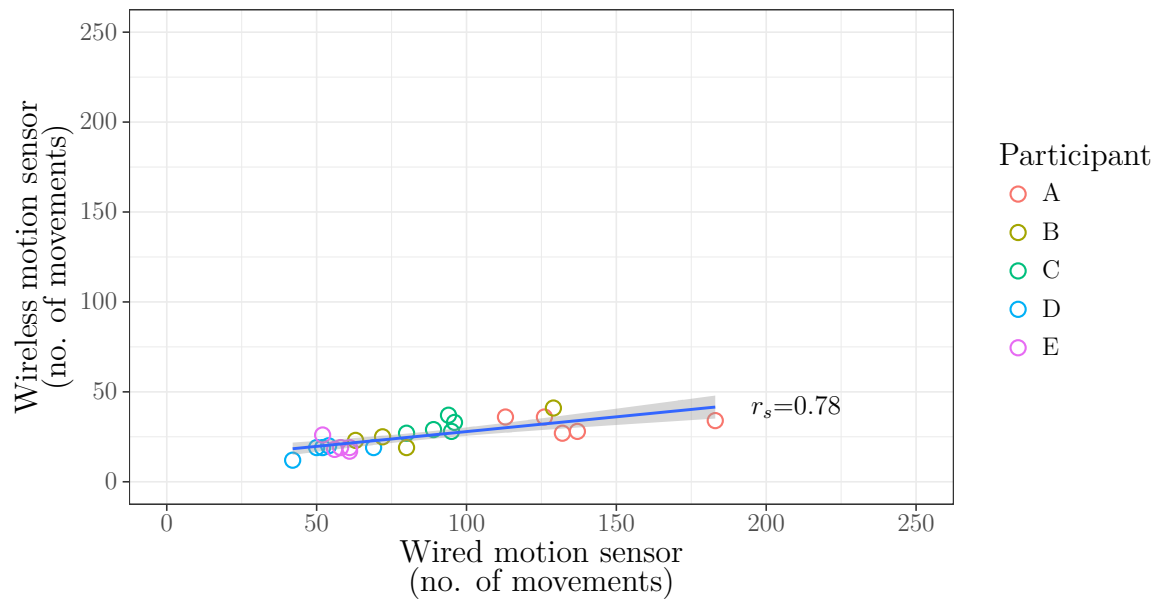


Figure 3.7: Correlation between the counts of hand movements recorded by wired and wireless (ACC) motion sensors.

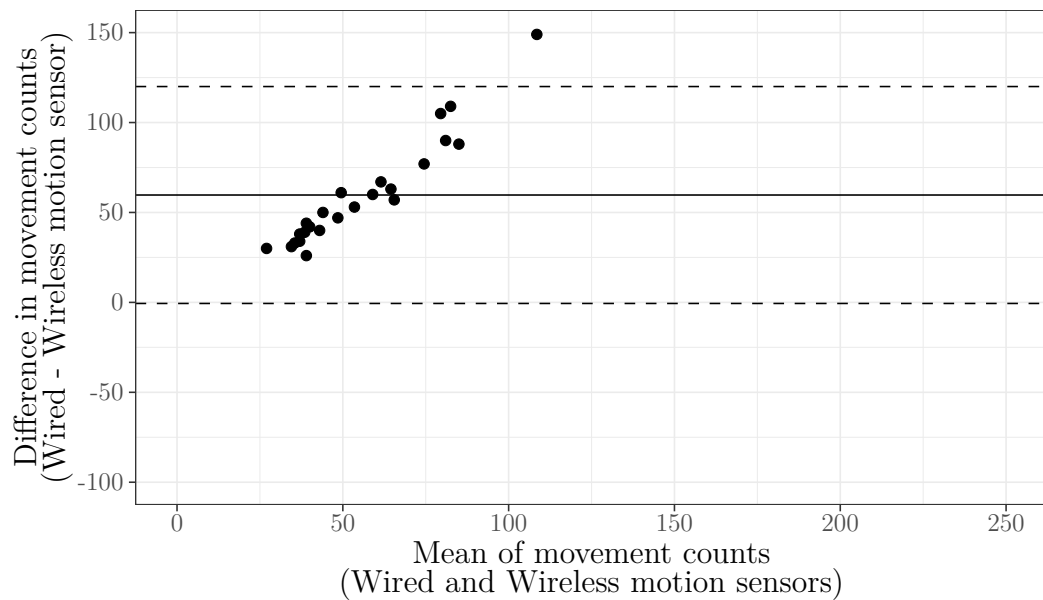


Figure 3.8: Agreement between the counts of hand movements recorded by wired and the wireless (ACC) system devices. Solid line: mean of differences in counts; dashed line: 95% limits of agreement (mean $\pm$ 2 S.D.) of differences in counts.

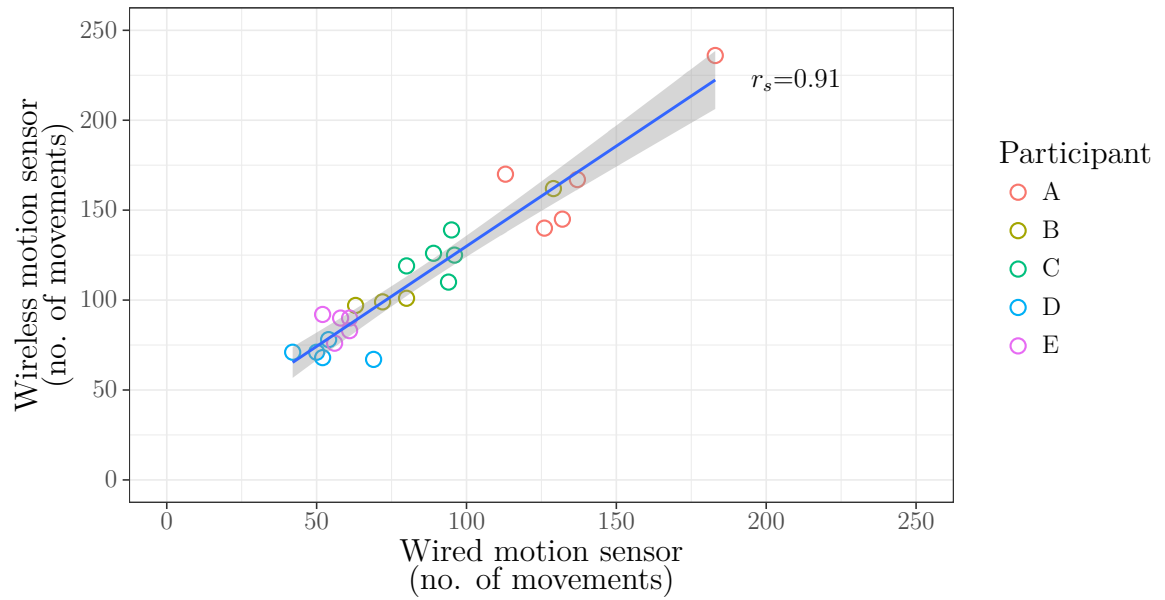


Figure 3.9: Correlation between the counts of hand movements recorded by wired and wireless(ACC+GYRO) system devices

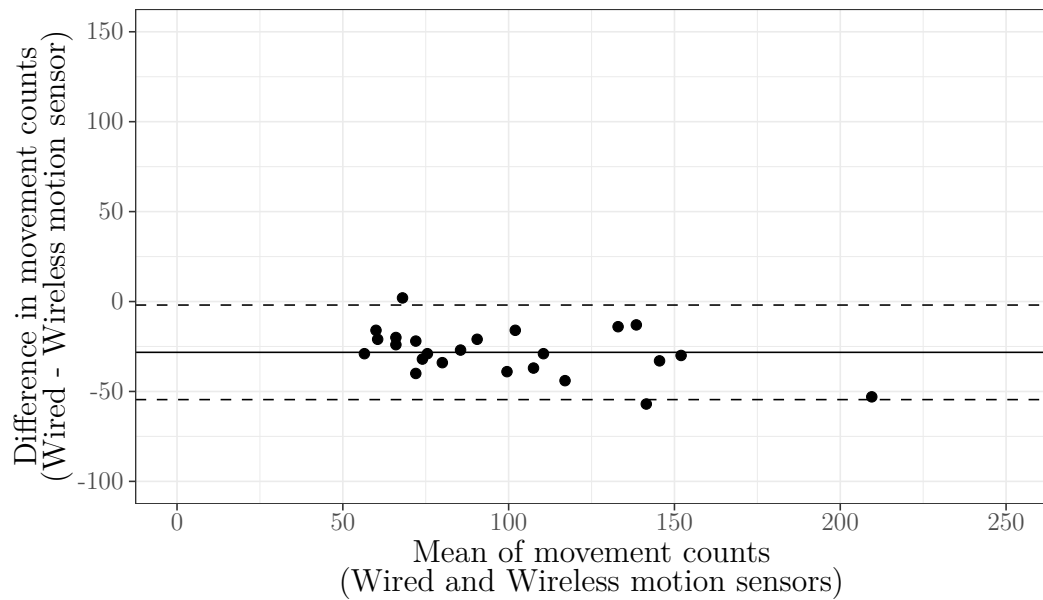


Figure 3.10: Agreement between the counts of hand movements recorded by wired and wireless(ACC+GYRO) system devices. Solid line: mean of differences in counts; dashed line: 95% limits of agreement (mean $\pm$ 2 S.D.) of differences in counts.

3.3 Study B - accuracy and reliability of wired and wireless motion analysis

The true number of movements required by each participant to complete the task in Study A was unknown, and as the simulator was novel to all participants it was anticipated there would be an improvement in performance with successive attempts. This makes the data from Study A unsuitable for investigating accuracy or reliability. Study B created an environment where the number of hand movements was controlled.

3.3.1 Methods

Ethical approval

The study was conducted under ethical approval from University of Oxford CUREC (MSD-IDREC-C1-2014-152; Appendix A.4, pg. 191).

Participants

The author was the only participant.

Study setting

Simulated arthroscopy was performed in the skills laboratory at OOSEC using a standard 30° arthroscope, Smith & Nephew 460H camera, light source, and image management system, with images displayed on a 15-inch Iiyama TFT monitor. The equipment and simulation station arrangement is described in Appendix A.

Arthroscopic task

A task was designed to which would allow a known number of movements to be performed. A rig with grooves to restrict movement of an arthroscopic probe to a fixed distance along orthogonal tracks in the x, y and z planes was designed and 3D-printed (Figure 3.11). Three rigs with 10, 5 and 1 cm long tracks were printed, to represent a range of movement magnitudes. The rigs were mounted in the centre of a FAST workstation (Fundamentals of Arthroscopic Training;

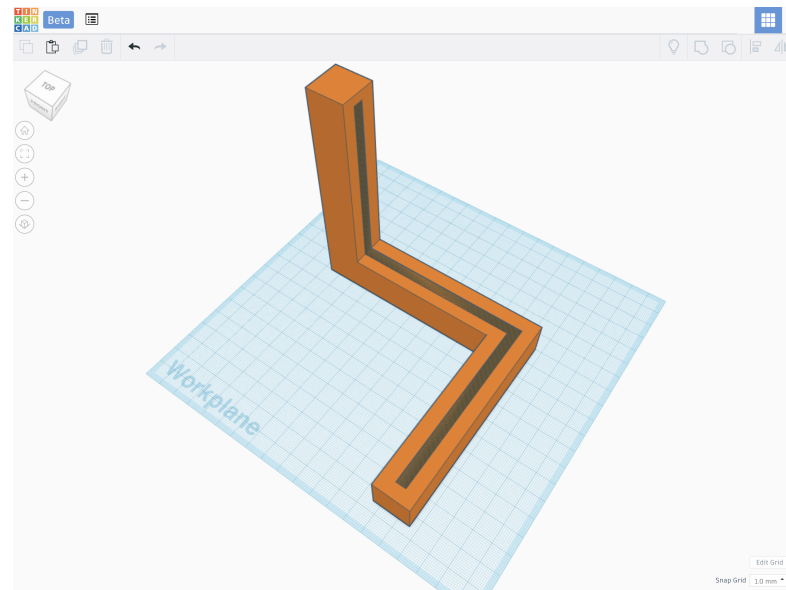


Figure 3.11: Design of 3D-printed rig with 10 cm long tracks used to restrict movements to generate a fixed number of stereotyped arthroscopic movements.

Sawbones, Malmö, Sweden), a generic arthroscopic skills training platform with interchangeable tasks (Figure 3.12). The task began with the participant at rest, holding the arthroscope in their right hand, the probe in their left hand, placed in predetermined portals on the FAST workstation. The probe tip was placed at the nearest end of the groove running in the z-axis. Once motion recording was started on both wired and wireless systems, the participant made 20 movements sequentially along each arm of each track in turn (z,x,y,x,z...). Once completed, the participant remained still until the motion recording had been stopped. The participant completed 10 attempts in total, three attempts at the 10cm rig, three at the 5cm rig, and four at the 1cm rig, while wearing both wrist-worn wired motion sensors and elbow-worn wireless motion sensors.

Motion analysis devices and metrics

The count of hand movements reported by the wired and wireless motions devices was recorded as described above (see Study A: Methods, pg. 80). Data were recorded from the left arm only, which was operating the probe. The other two motion sensors were placed flat on the table and not disturbed to ensure they didn't count any



Figure 3.12: External view of the FAST workstation generic arthroscopy simulator. The tasks inside are modular and can be exchanged.

movements. As in Study A, counts were calculated from the wireless system by two methods: wireless (ACC) using accelerometers alone, and wireless (ACC+GYRO) using both accelerometers and gyroscopes together.

Statistical analysis

Median and inter-quartile range were calculated for the wired and wireless systems counts of hand movements. Both the wireless (ACC) and wireless (ACC+GYRO) system were compared to the wired system using a two-sided Wilcoxon signed-rank test, with $p \leq 0.05$ taken to be statistically significant. All statistical analyses were performed in R v3.3.2.²⁸⁴

3.3.2 Results

After 10 attempts at causing 20 stereotyped movements of 1, 5 or 10cm*, the wired system median count of movements was 21 (IQR 18, 22; see Figure 3.13). The wireless (ACC) system tends to under-estimate the count of hand movements (median=13; IQR 4, 14) which was statistically significantly different from the wired system ($p=0.006$). The wireless (ACC+GYRO) system tends to slightly

*Four attempts at 1cm track, and three attempts on each of the 5 and 10cm tracks

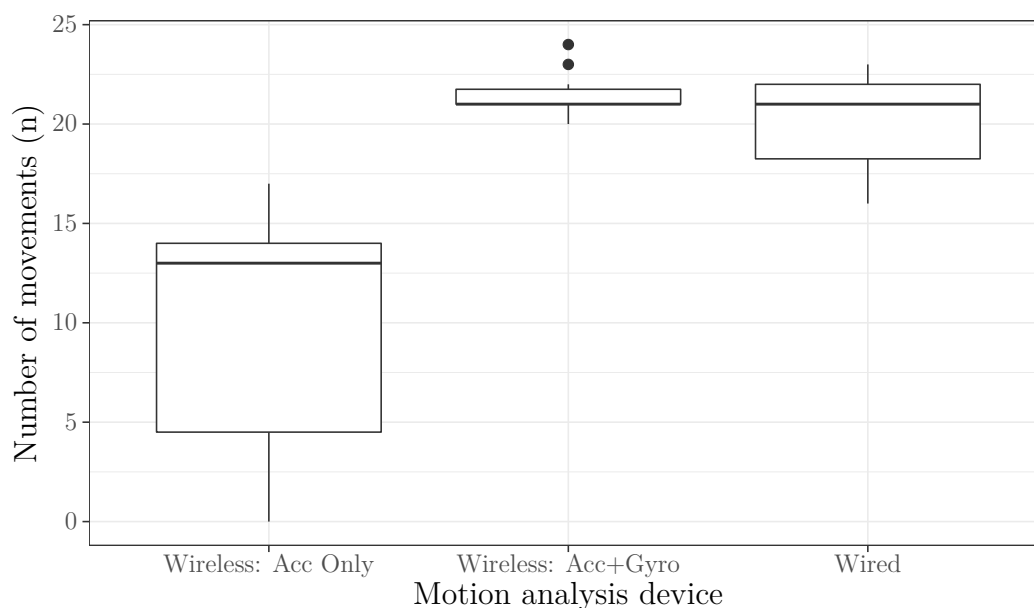


Figure 3.13: Reliability of the counts of hand movements reported by wired and both wireless motion analysis systems after 20 stereotyped arthroscopic movements.

over-estimate the count of movements (median = 21; IQR 21, 22) but this was not statistically significantly differently to the wired count ($p=0.117$).

Although not an *a priori* outcome of interest, it was noted that when inspecting the counts recorded on each of the rigs, the wireless (ACC) system counts relatively fewer movements when the track length is smaller compared with the wired and wireless (ACC+GYRO) system (Figure 3.14).

3.4 Discussion

3.4.1 Implications of the studies

These two studies indicate that of the two methods of counting hand movements from the WASP elbow-worn wireless motion analysis system, the method using both accelerometer and gyroscopic data is more reliable than using accelerometer data alone, and has a stronger correlation and better agreement with the Patriot wrist-worn wired motion analysis system supporting their concurrent validity. The counts of hand movements from the wired system and wireless (ACC+GYRO) system

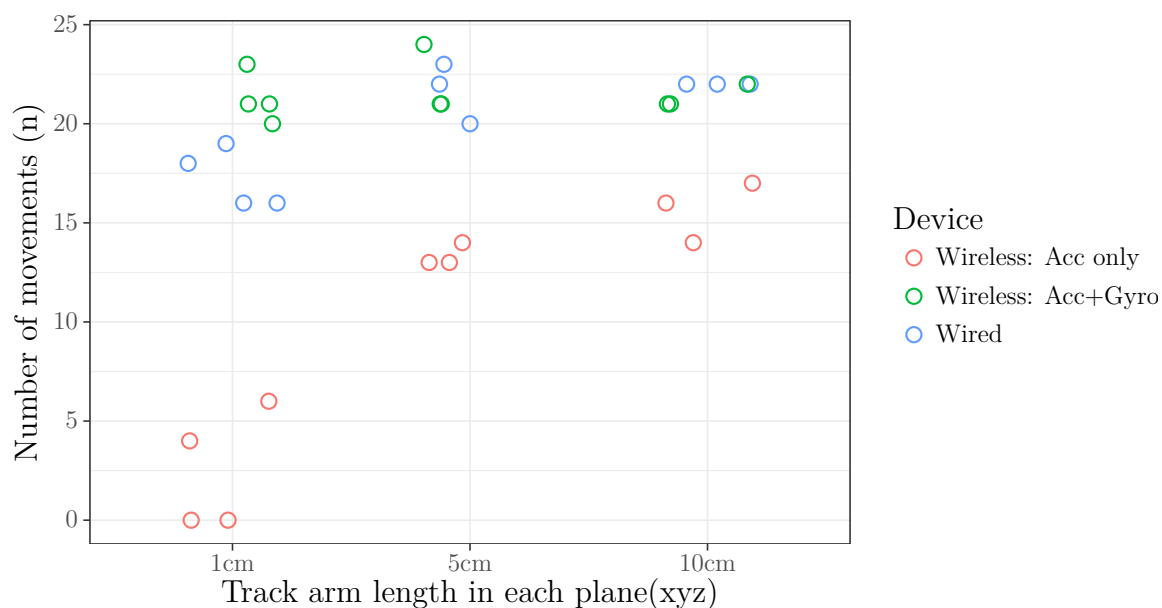


Figure 3.14: Reliability of the counts of hand movements compared by track length after 20 stereotyped arthroscopic movements

data are narrowly distributed and have medians similar to the intended number of movements performed, indicating their accuracy and supporting their face validity.

Most simulated and surgical tasks have an unknowable true of hand movements to test reliability and accuracy. Some virtual reality simulators calculate an optimal path length for a given task to present performance as a ratio of participant path length to optimal path length,^{187,233,238,239} but this is not possible beyond these specific VR tasks, and performance variation with practice over repeated attempts on these tasks would impair reliability and accuracy assessment. The rigs designed to restrict the direction and magnitude of movements in Study B help remove variability found within typical simulations or surgery. This allows the number of intended movements to be controlled. Stereotyped movements have been used previously in the simulation literature to calibrate motion analysis devices²⁷⁶ but this is the first study to combine ‘natural’ simulated tasks to assess the concurrence of two performance measures with stereotyped simulated tasks to assess reliability and accuracy.

When the count of hand movements reported by a motion analysis device is well-matched to the number planned by the user, this is an important mode of feedback which increases user's confidence that the system represents their real world activity. Accuracy and reliability allows users to be confident that the metric represents real world activity, which is an element of face validity.⁹⁸ In Study B, the wired and wireless (ACC+GYRO) system counts of hand movements were similar to the intended number of movements (both median 21 recorded movements cf. 20 intended movements), and the narrow interquartile range (wired IQR 18, 22; wireless (ACC+GYRO) IQR 21, 22) indicates the systems are reliable[†]. In contrast, the wireless (ACC) system counted consistently below the intended number of movements (median = 13) with a wider interquartile range (IQR 4, 14), indicating this method was less accurate and reliable.

This may seem at odds with the findings of study A, where there is a high positive correlation (Figure 3.7; $r_s = 0.78$) between the hand movement counts from wired system and wireless (ACC) system, if one is an accurate, reliable measure and the other is inaccurate and unreliable. However, correlation coefficients are sensitive to the range of the measures being compared,²⁸⁵ which is increased by the varied experience of the participants in study A (see Table 3.1). Bland-Altman plots show the degree of agreement between two measures, and the wireless ACC) system consistently underestimates the number of hand movements (mean difference in counts: 59.7). Moreover, this bias appears to increase with the count of hand movements (Figure 3.8). This may be the apparent insensitivity of the wireless (ACC) system to counting hand movements seen in study B, giving an unrealistically low impression of surgical activity and reducing confidence in the metric which undermines its face validity. By contrast, the wireless (ACC+GYRO) system with accelerometer and gyroscope data has a higher positive correlation with the wired system (Figure 3.9; $r_s = 0.91$), but it seems to overestimate the count of hand movements. The difference was not statistically significant over 20 movements

[†]It is unlikely the rig would suppress all extraneous movement, the overestimate of the wireless (ACC+GYRO) system compared to the intended 20 movements may in fact be closer to the true number of hand movements

in Study B, but during the longer tasks in Study A there was a consistent bias that the wireless (ACC+GYRO) system overestimated the counts (mean difference in counts -28.25; see Figure 3.10).

Why would the three methods have such variation in their sensitivity to detecting movements, especially given both wireless methods use the same accelerometers? The higher sampling frequency of the wireless system may detect transient accelerations which would be smoothed by the lower sampling frequency of the wired system, but this does not explain why the counts of the two wireless methods diverge from the wired method. There are 6 degrees of freedom for a rigid body in 3D space, three translational (surge, heave, sway) and three rotational (pitch, roll, yaw) axes, and although the final result is presented as hand movement count in all three methods it is inferred differently by each device:

- The wireless (ACC) system with only accelerometer data is only sensitive to the three translation movements
- The wired system records position and then integrates to get acceleration data used to count movements: this is sensitive to three translational axes, and receives surrogate information about rotation because the sensor on the wrist is offset from the centre of rotation and thus detects pronation and supination. By measuring position rather than acceleration itself, the wired system can only be sensitive to the net effect of acceleration in all six axes.
- The wireless (ACC+GYRO) system with accelerometer and gyroscope data measures acceleration in all six axes directly with three accelerometers and three gyroscopes. In practice, complex human movements with translational and rotational acceleration during arthroscopy which result in no absolute positional change to the sensor may contribute to the difference in each device's sensitivity to movements.

The wireless (ACC+GYRO) system agrees with the wired system better than the wireless (ACC) system, with a narrower distribution of differences between

measurements with the wired system (wireless (ACC+GYRO) s.d. 13.1 c.f. wireless (ACC) s.d. 30.2; see Figures 3.8 and 3.10). There is also a positive trend between the mean count of the wired and wireless (ACC) system and the difference in counts (Figure 3.8), not seen between the wired and wireless (ACC+GYRO) system (Figure 3.10). For the purposes of this study, the level of agreement between the wired system and wireless (ACC+GYRO) system indicated sufficient accuracy, reliability, and concurrent validity to consider further validation of the WASP system.

The studies above ask whether wireless elbow-worn devices could be used in place of commonly used wired wrist-worn devices to count hand movements during simulation, and ultimately be considered for live surgery as their position allows them to be worn during hand-washing. This wireless system has been used only once previously in the simulation literature, where the wireless devices were shown to give equivalent results when placed at the wrist and elbow, where the count of hand movements recorded correlated well with the path length and global rating scale scores recorded on a selection of virtual-reality tasks during a construct validation study.²²³ Although the VR simulator used in that study does record the count of hand movements also, there was no like-for-like comparison for agreement of equivalent metrics calculated by VR and wireless system, as in the studies A and B above.

Other methods of motion analysis in theatre should be considered. A precursor to the Patriot wired system has been used in porcine laparoscopic cholecystectomy surgery:²⁸⁶ all though these were real animals the procedure was for simulation purpose, and wired motion systems have not been used in therapeutic human interventions. Other wireless techniques are possible, for example, wireless sensors are available for the Patriot system,²⁸⁷ but these are larger and heavier than the WASP sensors²⁸⁸ ($42 \times 89 \times 25mm$ vs $23 \times 32.5 \times 8.9mm$, and 79g vs 8g), are untested in surgical simulation, and would still require the use of an electromagnetic field generator close to the patient in theatre which would be a regulatory obstacle. Surgical navigation systems could be co-opted to motion tracking in surgical simulation and theatre, but again this is untested.²⁸⁹ Mechanical systems to measure force could be made wireless and embedded in instruments or ‘force gloves’, but

these approaches have not been developed and pose regulatory and sterility issues respectively.⁷¹ An optical motion-capture system would be truly wireless with no need for sensors on the surgeon or instruments, but they have a lower static accuracy than magnetic and inertial systems and require a clear line-of-sight.⁸⁷ This can be controlled in the simulation laboratory but is more challenging in theatre.

3.4.2 Limitations

There is only one motion metric the wired and wireless motion analysis systems have in common, and the concurrent validity of the count of hand movements between the devices doesn't consider the other metrics which can be calculated: these will need to be evaluated separately. Each system calculates the number of hand movements differently from the raw data they collect (position in Patriot vs acceleration in WASP system; see Figure 3.2), and although in principal the wired systems position data might be coerced into the wireless system's algorithms it is impractical and beyond the scope of this thesis to interrogate the algorithms used to generate counts of hand movements. The wired system also reports path length, a validated measure of surgical dexterity (see Chapter 2): position data (which are used to calculate path length in the wired system) are differentiated to give acceleration data (which is used to count movements in both systems), but the reverse is not possible to allow the WASP system to calculate path length from acceleration data. It would be exquisitely sensitive to errors in starting conditions: with 1° error in starting orientation, after 10 minutes the position calculation would have an error of approx. 62km,²⁸² which is clearly impractical when calculating path length.

In these studies, accuracy and reliability has not been assessed in a 'real' task. The 20 movement rig is a surrogate, and represents very artificial stereotyped sequential movements restricted in their plane and magnitude. The pretence permits a contrived assessment of accuracy and reliability, but the findings above may not apply to natural human movements. In 'real' tasks, the unknown number of movements and the effect of learning and improvement by the participants over repeated attempts preclude assessment of accuracy and reliability, respectively.

3.4.3 Summary

These studies indicate that a wireless (ACC+GYRO) system using both accelerometer and gyroscope data is a reliable and accurate measure of movement during simulated arthroscopy. Their methods of calculating the number of hand movements correlates well with a commonly used wired system with positional motion sensors, and agreement is sufficient to justify further validation of this wireless system. The concurrent validity, accuracy and reliability of wirelessly acquired counts of hand movements have been supported, but validation of the other metrics calculated by the system need to be evaluated, as do other aspects of validity for the proposed inferences in using the same motion sensor metrics in a simulated and real surgical setting. A stated aim of this thesis is to objectively validate the transfer of simulation training into the operating theatre, and for this purpose the WASP system was selected as it was cheap, required little additional equipment (two sensors and a laptop), and appears to more suitable for arthroscopic assessment than other solutions.

It's not the simulator. It's the curriculum.

— Richard Satava

4

Development of a simulated arthroscopy list

Contents

4.1	Introduction	97
4.1.1	Aim	99
4.2	Study A - Task selection for a simulated arthroscopy list	100
4.2.1	Methods	100
4.2.2	Results	105
4.3	Study B - Construct and face validity of a simulated arthroscopy list	108
4.3.1	Methods	108
4.3.2	Results	111
4.4	Discussion	122
4.4.1	Implications of the studies	122
4.4.2	Limitations	125
4.4.3	Summary	126

4.1 Introduction

Deliberate practice relies on accurate performance feedback at the time of the task to guide progress and is key to achieving expertise in surgery and other high-performance disciplines.¹⁰⁰ A distinction is made between the experience gained

from activities of ‘work’ (“public performance, competitions, services rendered for pay, and other activities directly motivated by external rewards”) and ‘deliberate practice’ (“activities that have been specifically designed to improve current level of performance”).⁹³ Routine surgical exposure in the traditional model of surgical training provides experience but falls entirely within ‘work’: weaknesses in ability may be exposed and exhibit their consequences, but cannot be systematically explored, tested repeatedly, and refined. However, many surgical procedures and techniques have generalisable components that could be isolated and practised in a simulated setting,⁹⁹ and arthroscopy has particular psychomotor skills amenable to simulated practice.⁹⁸ The Fundamentals of Arthroscopic Surgery Training (FAST) workstation has been developed by the Arthroscopy Association of North America (AANA) as generic arthroscopic skills trainer with eight interchangeable training tasks to allow deliberate practice,²⁹⁰ but relies heavily on time taken to measure performance which is a poor metric as discussed in Chapter 2. Other metrics have been validated to assess performance but not whether they are an effective feedback method for training: even when used in proficiency-based training, trainees received their feedback from a supervisor^{129,291} rather than undertaking independent deliberate practice being guided by their metric performance.

Construct validity describes whether a task measures what we intend it to measure and is an feature of our inferences based on the task itself, how performance is measured, and the context to which it is applied.^{159,160} In surgical simulation, we use participants with expected differences in performance due to their surgical experience as a standard to test how a recorded metric during a task differentiates between participants. If the performance results can distinguish those with expected or known performance differences, the inference is that the inverse holds true: performance differences between participants with unknown expertise represent differences in real world performance. In training and practice scenarios, deliberate practice on a variety of construct validated tasks with feedback identifies areas of performance weakness and permits knowledge of results to guide progress with the expectation improvements would be reflected in real surgery.¹⁰³

The goals of deliberate practice are important, and trainees are informed of these by their trainers and training programme. The Trauma and Orthopaedic (T&O) curriculum sets out the level of competence expected in various procedures in order to receive a certificate of completion of training (CCT).⁸⁹ Competence is assessed by logbook entries (at least 40 arthroscopic cases performed), and requirements for outcomes in procedure-based assessments (PBA) and case-based discussions (CbD) with their trainers. For arthroscopic meniscectomy, partial meniscectomy, and diagnostic knee arthroscopy, trainees are required to achieve the highest level of competence as assessed by an external observer.²⁹²

To improve the implementation of simulation in orthopaedics requires a validated training environment for the deliberate practice of specific arthroscopic skills using meaningful objective feedback from motion performance metrics to inform areas of weakness and improvements with practice, which is consistent with the final goals of the curriculum.

4.1.1 Aim

The aims of this chapter are to develop an environment (termed the ‘simulated arthroscopy list’, or SAL) where the essential arthroscopic skills can be deliberately practiced, and to validate that environment and the use of motion metrics to measure performance. This was achieved by two linked studies (Figure 4.1):

Study A to select the FAST tasks most appropriate for training novice arthroscopists in the skills required to reach the competencies described by the T&O curriculum.

Study B to assess the construct and face validity of a simulated arthroscopy list composed of i) the selected tasks in Study A, ii) a novel meniscectomy model, iii) and a validated diagnostic knee arthroscopy model, used with the motion metrics produced by the WASP system

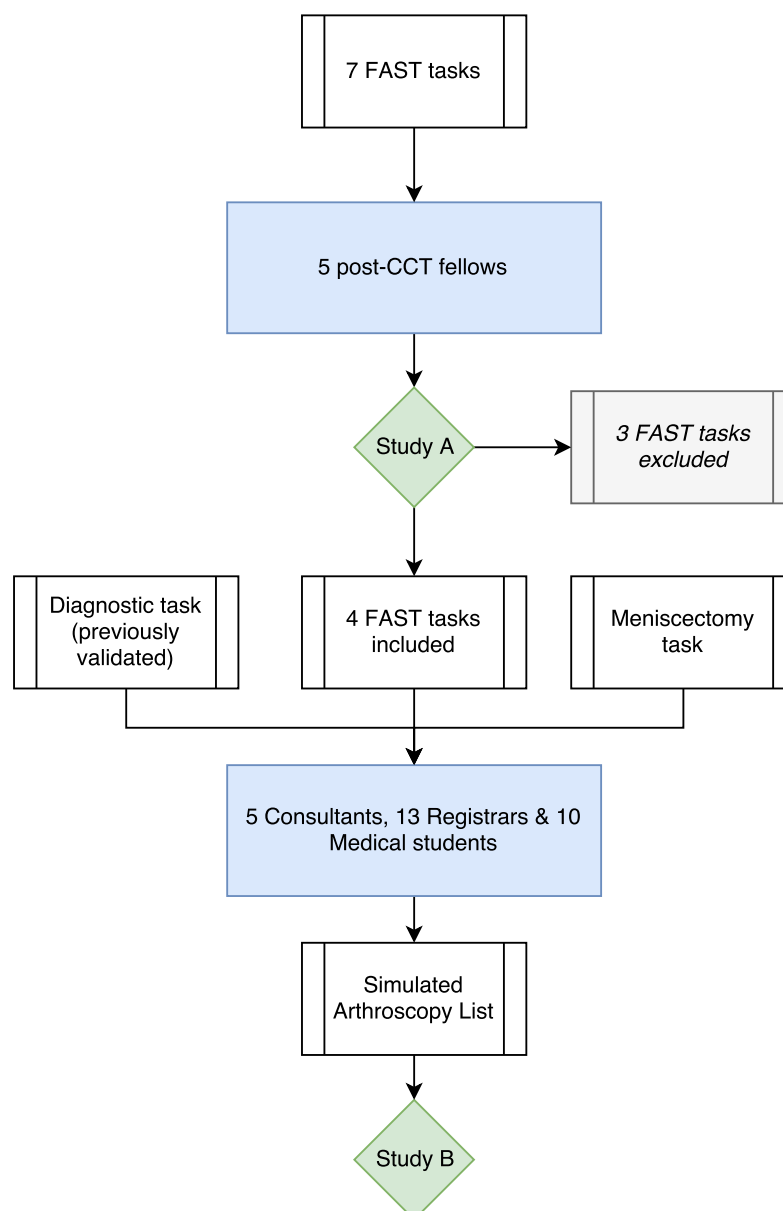


Figure 4.1: The relationships between the studies conducted to develop the simulated arthroscopy list

4.2 Study A - Task selection for a simulated arthroscopy list

4.2.1 Methods

Ethical approval

The study was conducted under ethical approval from University of Oxford CUREC (MSD-IDREC-C1-2014-152; Appendix A.4, pg. 191).

Participants

Five arthroscopy fellows were invited to participate to provide a consensus view on the most appropriate tasks from the perspective of someone with significant arthroscopic experience who has recently completed the T&O training programme. The fellows were all employed in post-CCT fellowships, and had recently achieved arthroscopic competence for the purpose of CCT (at least 40 arthroscopies performed; level 4 PBA and CbD) but were yet to take up a consultant post.

Study setting

Simulated arthroscopy was performed in the skills laboratory at OOSEC using a standard 30° arthroscope, Smith & Nephew 460H camera, light source, and image management system, with images displayed on a 15-inch Iiyama TFT monitor. The equipment and simulation station arrangement is described in Appendix A.

Fundamentals of Arthroscopic Surgery Training (FAST) tasks

The FAST workstation (#1711; Sawbones, Malmö, Sweden) is a bench-top generic arthroscopic skills training system with a base which accepts different task modules to practice arthroscopic skills, covered by either a clear or opaque plastic dome with predefined arthroscopic portals. The available modules are shown in Figure 4.2:

Probing Participants find and probe 21 randomly arranged numbers in the correct numerical order

Maze Participants navigate a ball bearing around a maze using a probe. If the ball bearing is knocked over the walls to another part of the track, it must be returned to its previous position before continuing

Flat ring transfer Participants remove eight green rings from metal pegs pointing vertically from a horizontal plane and place them in the centre well. Once the first stage is completed, return the green rings to vertical metal pegs on the opposite side of the trainer. The number of dropped rings was recorded

Column ring transfer As for the flat ring transfer, but the metal pegs are arranged in a 3x3 arrangement horizontally attached to a cylindrical column

Arc cutting Participants use meniscal punches to cut a semi-circular pattern between two black lines on a silicone disc

Knot tying Participants tie an arthroscopic knot of their choosing around a metal post

Cuff repair Participants insert a bone anchor, pass the sutures through simulated cuff tissue, then secure it with an arthroscopic knot

Smartphone holder (Not shown) This module was not used in this study, as at that time no smartphone applications had been developed for the practice of arthroscopic skills

Curriculum

The participants were all reminded of the contents of their T&O training curriculum and shown how the knee arthroscopy tasks were mapped to competency levels at different stages of training. In particular, the ‘Arthroscopy knee diagnostic task’, ‘Arthroscopic meniscectomy’ and ‘Arthroscopic partial meniscectomy’ tasks were highlighted as these represented three tasks where the highest level of competency (5: “able to manage complex cases and their potential associated complications”²⁹²) is expected at the end of training for all orthopaedic surgeons (Table 4.1).

Assessments

Each of the participants gave informed consent before watching an orientation PowerPoint presentation demonstrating how each task should be performed. The participants completed a single attempt at each of the FAST tasks, in approximately increasing order of difficulty as laid out by the Arthroscopy Association of North American (AANA) and American Board of Orthopaedic Surgery (ABOS).²⁹³ After attempting the tasks, they completed a ‘FAST task selection’ survey online using Google Forms (Figure 4.3). The survey asked participants to: rate the complexity of each task; rank the tasks based on their usefulness for novice trainees; suggest which tasks they would select for novice trainees to practice the skills required by the curriculum; and whether they felt the two ring transfer tasks were sufficiently



Figure 4.2: FAST arthroscopy workstation tasks used in this study. Clockwise from top-left: external appearance of FAST trainer, Probing, Flat ring transfer, Meniscal cutting, Cuff repair, Knot tying, Column ring transfer, and Maze

Table 4.1: T&O curriculum knee arthroscopy procedures mapped to expected competencies at each stage of training²⁹²

Procedure	Core	ST3-8	Specialty
Acute arthroscopy for knee trauma	0	3	5
Irrigation & debridement native joint for infection - knee	1	4	5
Irrigation & debridement prosthesis for infection - knee	1	4	5
Acute ligament repair	0	3	5
Tibial plateau fracture arthroscopically assisted fixation	0	2	5
Arthroscopic partial meniscectomy	1	5	5
Arthroscopic excision of Hoffa's fat pad	0	4	5
Arthroscopic lateral release	0	4	5
Arthroscopic meniscectomy	0	5	5
Arthroscopic removal loose bodies knee	0	4	5
Arthroscopic synovectomy	0	3	5
Arthroscopy knee diagnostic	1	5	5
Meniscal repair (arthroscopic)	0	3	5
Abrasion arthroplasty / microfracture - knee	0	2	5
Mosaicplasty - knee	0	2	4
Osteochondral allografting - knee	0	2	4
Autologous chondrocyte implantation	0	2	5
ACL reconstruction	0	2	5
Reconstruction of PLC	0	2	5
PCL reconstruction	0	2	5
Revision ACL reconstruction	0	1	5

ST: Specialist training year. Skill levels as per 2014 curriculum: 0 - No experience expected; 1 - Has observed or knows of; 2 - Can manage with assistance; 3 - Can manage whole but may need assistance; 4 - Able to manage without assistance including potential common complications; 5 - Able to manage complex cases and their associated potential complications.

different, or developed separate skills, that they would include both in a training programme. Survey questions can be viewed in Appendix C.1, or online at tinyurl.com/PGR-thesis1.

As well as evaluating the FAST tasks for their suitability for training, the participants' performance would contribute to the overall validity assessment of the SAL in Study B. Their performance throughout was recorded using the WASP wireless elbow-worn motion analysis system, and they also attempted the diagnostic knee arthroscopy and arthroscopic partial meniscectomy exercises after the FAST tasks (see Study B, below). Their performance on any FAST tasks which were eventually excluded was ignored.

FAST Task Selection

All responses are confidential, and not used to assess you.

***Required**

How would you rate the complexity of the tasks you have performed? *

	Basic	Intermediate	Advanced
Task 1 - Probe	☐	☐	☐
Task 2 - Maze	☐	☐	☐
Task 3 - Flat ring transfer	☐	☐	☐
Task 4 - Column ring transfer	☐	☐	☐
Task 5 - Arc cutting	☐	☐	☐
Task 6 - Knot tying	☐	☐	☐
Task 7 - Cuff repair	☐	☐	☐

Figure 4.3: Excerpt from online FAST Task Selection survey available at tinyurl.com/PGR-thesis1

Outcomes and statistical analysis

Descriptive analysis of the survey would be used to establish a consensus on the FAST tasks to include in the simulated arthroscopy list. Performance on the tasks was not planned to contribute to the decision of whether or not to include a given task, though motion analysis performance data from the attempts would be collected for the construct validity assessment in Study B. All statistical analyses were performed in R v3.3.2.²⁸⁴

4.2.2 Results

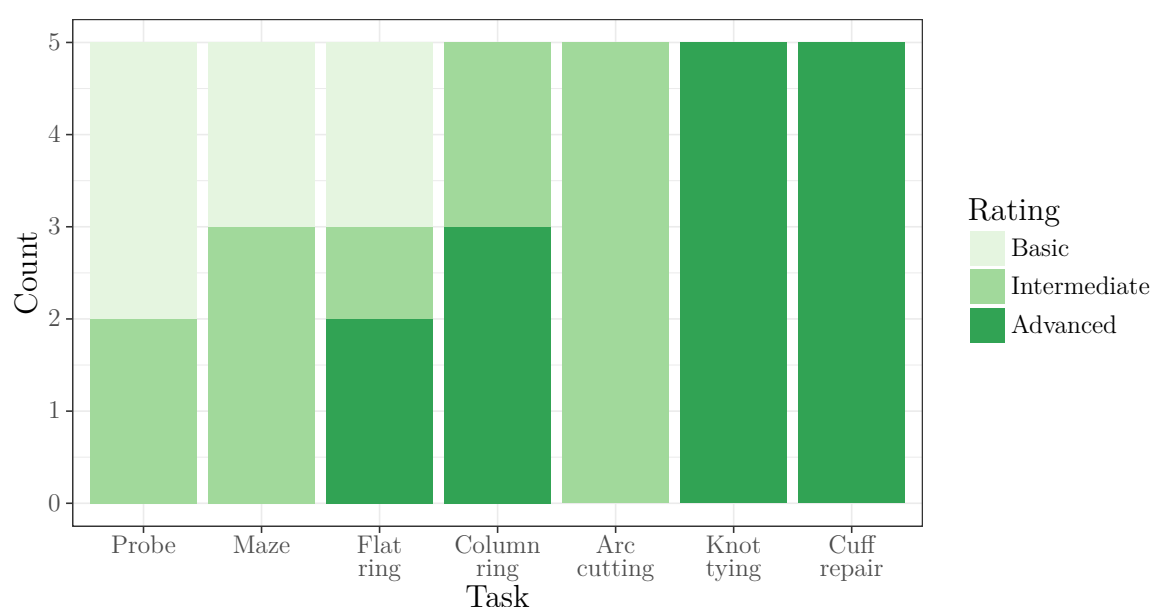
All five fellows accepted the invitation to participate. The fellows had a mixture of upper and lower limb interest. Their arthroscopic experience ranged from 72 to 293 cases performed, and all participants had previous exposure to surgical simulation during their training (Table 4.2).

Table 4.2: Basic characteristics and experience of the post-CCT fellows

Participant	Subspecialty	Arthroscopies performed	Handedness	Simulator exposure
A	Knee	72	Right	Bench-top
B	Hip	293	Right	Bench-top, Cadaveric
C	Shoulder	260	Right	Bench-top, Cadaveric, VR
D	Hip/Knee	270	Right	Bench-top, VR
E	Hip/Knee	172	Right	Bench-top, VR

Difficulty rating

The difficulty ratings for each task are summarised in Figure 4.4. The Knot tying and Cuff repair tasks were unanimously felt to be ‘Advanced’ skills*. The Arc cutting task was unanimously felt to be an ‘Intermediate’ skill, and the Probe and Maze tasks were either ‘Basic’ or ‘Intermediate’. The ring transfer tasks had a mixture of responses, with the Column ring transfer felt to be more advanced.

**Figure 4.4:** Post-CCT fellows rating of the difficulty of each task.

*In fact, only the participant with a shoulder subspecialty interest was able to complete these two tasks.

Usefulness rating

The usefulness ratings are summarised in Figure 4.5. The tasks felt to be more difficult (see Figure 4.4), tended to be rated as less useful for training novices. Each of the Probe, Maze, Flat ring transfer and Arc cutting had at least one participant rating them in the top three most useful exercises. The Column ring transfer, Knot tying and Cuff repair tasks were unanimously in the bottom three most useful exercises for novice trainees.

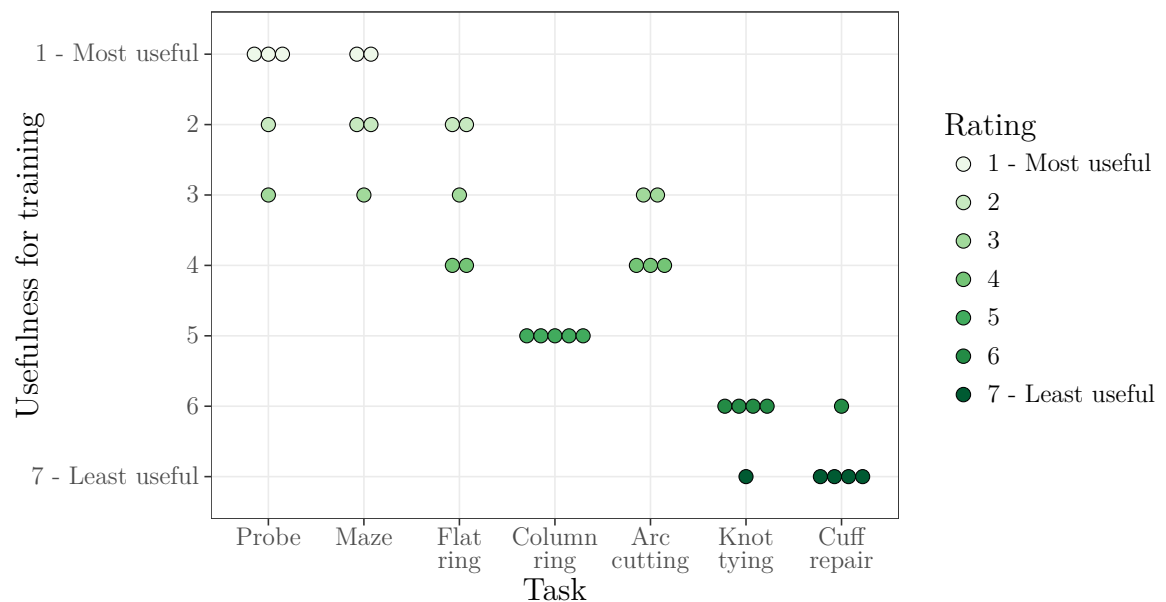


Figure 4.5: Post-CCT fellows rating of the relative usefulness of each task for novice trainees.

Suggested tasks for inclusion

All participants suggested the Flat ring transfer and Arc cutting tasks would be suitable to train the skills required for diagnostic and therapeutic arthroscopy to novice trainees. Four participants suggested the Probe task, and three suggested the Maze task. None suggested the Column ring transfer, Knot tying or Cuff repair tasks.

Difference between ring transfer tasks

Three of the five participants felt that the Flat and Column ring transfer tasks were sufficiently different, or developed separate skills, that they would consider including them separately in a training programme.

Interpretation and implications for Study B

The four tasks suggested for inclusion (Probing, Maze, Flat ring transfer, and Arc cutting) were those felt by the participants to be most useful skills and not too advanced for novices. Despite the majority of the participants feeling the two ring transfer tasks were sufficiently different, no participant felt that both should be included in the simulated arthroscopy list going forward. Thus, four modules were included in the simulated arthroscopy list to be validated in Study B.

4.3 Study B - Construct and face validity of a simulated arthroscopy list

4.3.1 Methods

Ethical approval

The study was conducted under ethical approval from University of Oxford CUREC (MSD-IDREC-C1-2014-152; Appendix A.4, pg. 191).

Participants

This study recruited three groups of participants with anticipated real differences in technical skill in line with the stages of the Dreyfus model:²⁴⁸ medical students with no arthroscopic experience (Novice group, approximating to “novice” stage), orthopaedic registrars (Intermediate group, approximating to “proficient” stage), and post-CCT orthopaedic fellows & consultants (CCT group, approximating to “mastery” stage). The CCT group were invited in person from staff at the Nuffield Orthopaedic Centre, and the Intermediate and Novice groups were invited by message sent out on departmental or medical school email lists. Eligibility for

each group was defined by Table 4.3 and participants were excluded if they had previous experience of the FAST trainer.

Table 4.3: Eligibility criteria for each experience level recruited to Study B

Group	Recruited from	Arthroscopic experience		
		Observed	Assisted	Performed
Novice	Medical students	<10	0	0
Intermediate	Orthopaedic registrars	-	-	10<x<40
CCT	Post-CCT fellows and consultants	-	-	>40

Study setting

Simulated arthroscopy was performed in the skills laboratory at OOSEC using a standard 30° arthroscope, Smith & Nephew 460H camera, light source, and image management system, with images displayed on a 15-inch Iiyama TFT monitor. The equipment and simulation station arrangement is described in Appendix A.

Simulated arthroscopy list tasks

This study included the four FAST modules identified in study A (Probing, Maze, Flat ring transfer, and Arc cutting), a novel meniscectomy task (see Figure 4.6 and Appendix C.2), and a previously validated diagnostic knee arthroscopy model as a reference task⁸⁴ (#1401; Sawbones, Malmö, Sweden). The FAST Flat ring module was recorded as two separate tasks in the WASP client: ‘Ring removal’, taking the green rings off the vertical metal peg and placing them in the central well, and ‘Ring return’, picking the green rings out of the central well and placing them back onto the vertical metal pegs. This was because during initial testing return of the rings was felt to be more challenging and ought to be recorded separately.

Assessments

Each of the participants gave informed consent before watching an orientation PowerPoint presentation demonstrating how each task should be performed. The participants completed a single attempt at each of the FAST tasks, in approximately increasing order of difficulty as laid out by the AANA and ABOS, then the diagnostic



(a) Diagnostic round: points to probe marked (b) Meniscectomy: tear locations marked

Figure 4.6: Anatomic knee arthroscopy simulated tasks included in the simulated arthroscopy list.

knee arthroscopy task and meniscectomy task, while wearing WASP motion sensors. After attempting the simulated arthroscopy list, they completed a face and content validity survey online using Google Forms. Survey questions can be viewed in Appendix C.3, or online at tinyurl.com/PGR-thesis2.

Outcomes

The primary outcome was WASP motion data about the following metrics which were hypothesised to be influenced by surgical experience: number of hand movements, number of minor hand movements, smoothness of movements, maximum acceleration, stationary time[†], inactivity time[‡], and difference in hand dominance[§].

The secondary outcome was the results of the survey which asked participants to comment on the following (see Appendix C.3 or online for full question stems and response items):

- How motivating and useful each of the metrics were
- Whether they knew how to interpret and improve their performance
- What impact they expected training on the FAST and meniscectomy tasks to have on arthroscopic skills⁹⁸

[†]The percentage of time where any hand is not moving

[‡]The percentage of time where *both* hands are not moving

[§]The percentage of time where either hand is more active than the other

- The realism of the meniscectomy task

Statistical analysis

Motion analysis data were analysed using non-parametric methods as the data were not expected to be normally distributed. Differences in performance at the all-groups level were assessed using the Kruskal-Wallis test for each metric on each of the tasks. Where statistically significant differences in performance were detected, group-pairs (Novice vs Intermediate, and Intermediate vs CCT) were compared using Mann-Whitney U test to see which groups with expected differences in performance the metrics could detect differences in. A p-value of <0.05 was taken to be significant, so for each metric multiple comparisons on the seven tasks were accounted for with a Bonferroni-corrected p-value of <0.0071 ($\frac{0.05}{7}$).²⁹⁴ Survey results were reported using descriptive statistics. All statistical analyses were performed in R v3.3.2.²⁸⁴

4.3.2 Results

Those who responded who were eligible to participate formed three similar sized groups with 10 post-CCT fellows/consultants (CCT group), 13 registrars (Intermediate group) and 10 medical students (Novice group). The CCT group had more surgical and simulator experience than the Intermediate and Novice groups (Table 4.4).

Time taken

The diagnostic knee arthroscopy task demonstrated significant performance differences as measured by time at the all-groups level and between each of the group-pairs ($p < 0.001$ in all cases), indicating that the CCT groups quicker than the Intermediate group, who were in turn quicker than the Novice group (Figure 4.7).

The Probe, Ring remove, Ring return, Arc cutting and Meniscectomy tasks all demonstrated significant performance differences at the all-groups level ($p < 0.001$ in all cases), which were also present at the group-pairs level between the Novice and Intermediate groups ($p \leq 0.0071$ in all cases). However, only the Probe and

Table 4.4: Baseline characteristics of participants in Study B.

		CCT	Intermediate	Novice
Participants		10	13	10
Arthroscopies	Median	1146.5	30	0
	IQR	216-2375	17-34	0
Right handed		100%	100%	90%
Simulation training	Daily	0%	0%	0%
	Weekly	0%	0%	0%
	Monthly	0%	0%	0%
	Termly	0%	0%	0%
	Annually	0%	15%	0%
	Less often	40%	54%	0%
	None	60%	31%	100%
Simulator experience	Animal	10%	0%	0%
	Benchtop – anatomic	50%	23%	0%
	Benchtop – non-anatomic	40%	38%	20%
	Cadaveric	60%	23%	0%
	Virtual Reality	30%	23%	0%
	None	20%	23%	80%

Meniscectomy task demonstrated a significant difference between the Intermediate and CCT groups ($p < 0.0071$ in both cases; Appendix C.1).

On the Maze task, performance as measured by time was similar for all three groups, with no statistically significant all-groups difference ($p = 0.52$).

Number of hand movements

The diagnostic knee arthroscopy task demonstrated significant performance differences as measured by number of hand movements at the all groups level and between each of the group-pairs ($p < 0.001$ in all cases), indicating that the CCT groups used fewer movements than the Intermediate group, who in turn used fewer movements than the Novice group to complete the task (Figure 4.8).

The Probe, Ring remove, Ring return, Arc cutting, and Meniscectomy tasks all demonstrated significant performance differences at the all-groups level ($p < 0.001$ in all cases), which were also present at the group-pairs level between the Novice and Intermediate groups ($p < 0.0071$ in all cases). However, only the Probe task demonstrated a significant difference between the Intermediate and CCT groups ($p = 0.00071$; Appendix C.2).

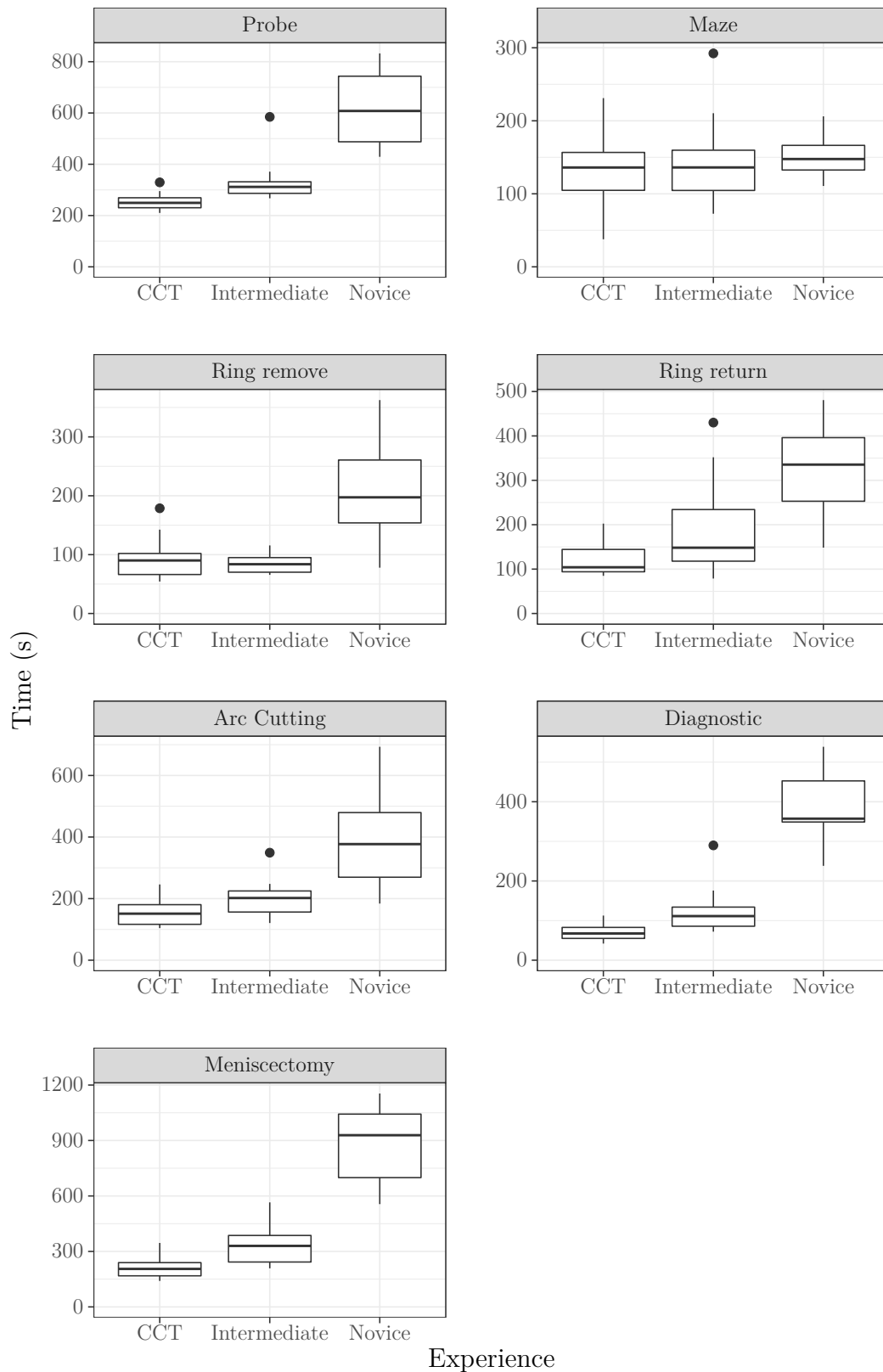


Figure 4.7: Performance on each simulated task measured by time taken (s). (Centre bar: median; Hinges: 1st & 3rd quartile; Whiskers: 1.5 x interquartile range; Points: outliers)

On the Maze task, performance as measured by number of hand movements was similar for all three groups, with no statistically significant all-groups difference ($p=0.041$).

Number of minor hand movements

The diagnostic knee arthroscopy task demonstrated significant performance differences as measured by number of minor hand movements at the all groups level and between each of the group-pairs ($p<0.005$ in all cases), indicating that the CCT groups used fewer minor movements than the Intermediate group, who in turn used fewer movements than the Novice group to complete the task (Figure 4.9).

The Probe, Ring remove, Arc cutting and Meniscectomy tasks all demonstrated significant performance differences at the all groups level ($p<0.005$ in all cases), which were present at the group-pairs level between the Novice and Intermediate groups in the Probe, Ring remove and Meniscectomy tasks ($p<0.0071$ in all cases). None of these tasks demonstrated a significant difference between the Intermediate and CCT groups (Appendix C.3).

On the Ring return and Maze tasks, the difference in number of minor hand movements was not statistically significant at the all-groups level ($p=0.012$ and $p=0.99$, respectively) when considering the Bonferroni corrected p-value.

Smoothness of movements

The diagnostic knee arthroscopy task demonstrated significant performance differences as measured by total smoothness at the all groups level and between each of the group-pairs ($p<0.005$ in all cases), indicating that the CCT groups were smoother (less jerky) than the Intermediate group, who were in turn smoother than the Novice group (Figure 4.10).

The Probe, Maze, Ring remove, Ring return, Arc cutting, and Meniscectomy tasks all demonstrated significant performance differences at the all groups level ($p<0.005$ in all cases), which were also present at the group-pairs level between the Novice and Intermediate groups ($p<0.005$ in all cases). None of these tasks

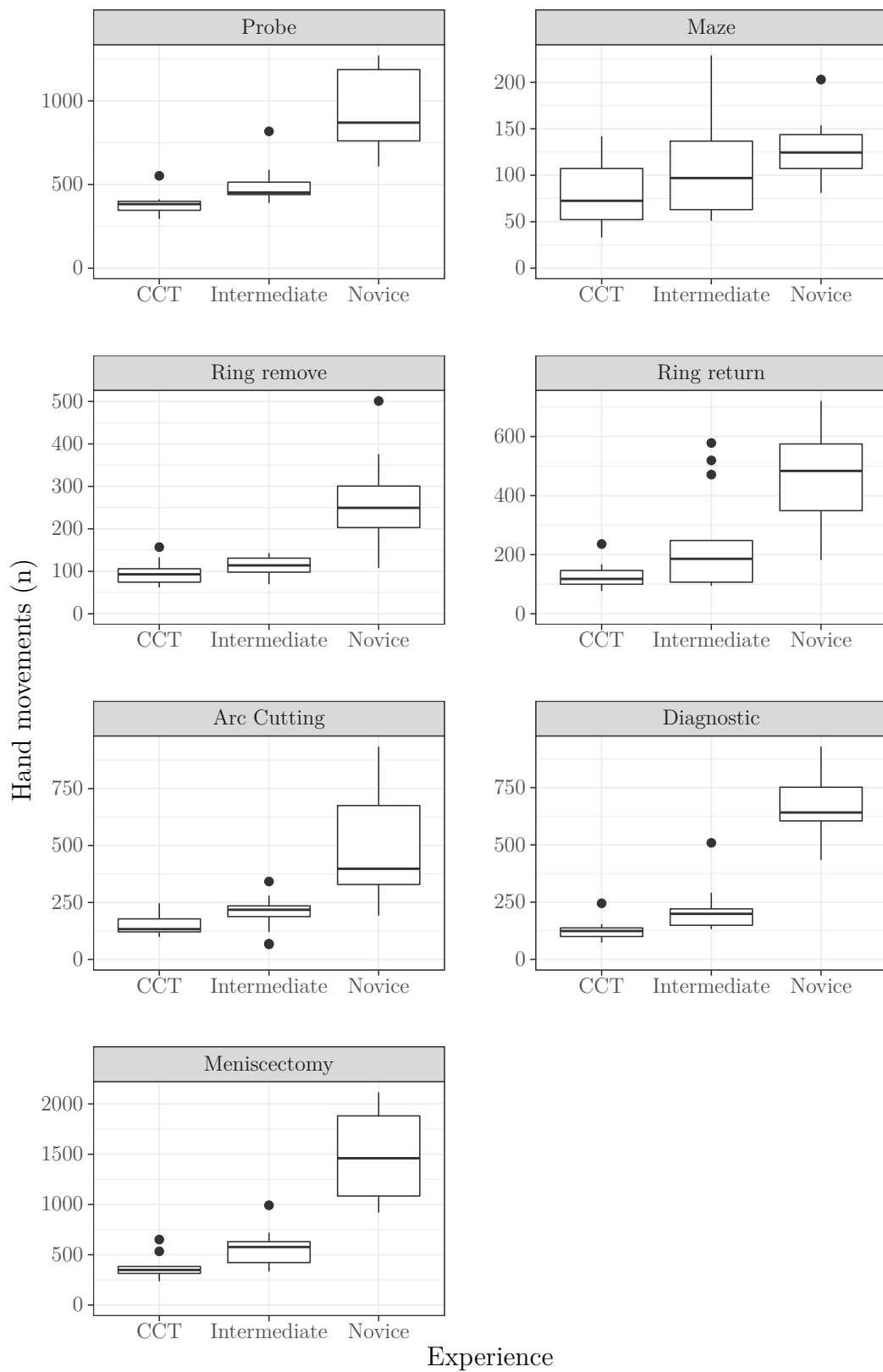


Figure 4.8: Performance on each simulated task measured by hand movements (n). (Centre bar: median; Hinges: 1st & 3rd quartile; Whiskers: 1.5 x interquartile range; Points: outliers)

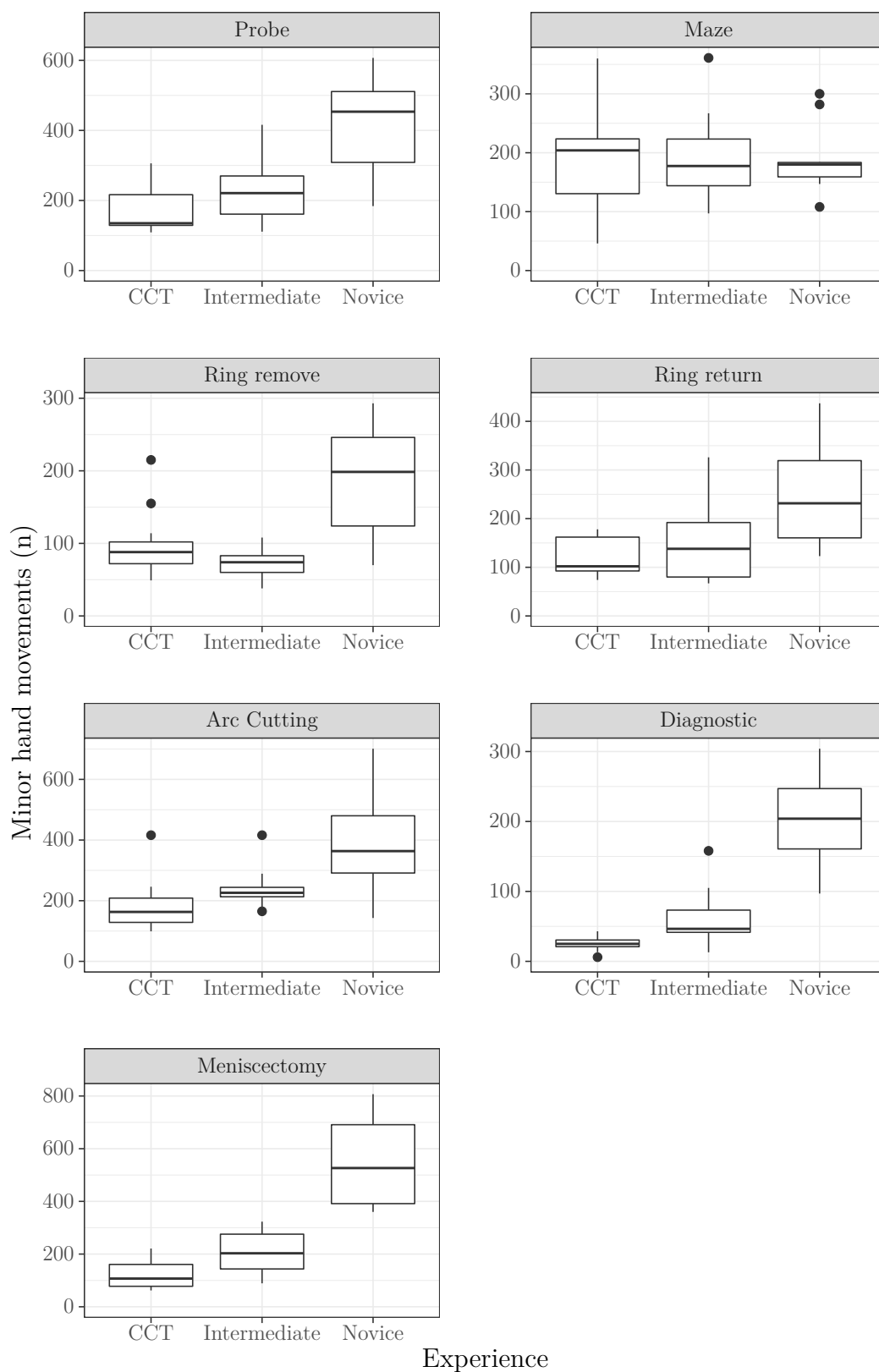


Figure 4.9: Performance on each simulated task measured by minor hand movements (n). (Centre bar: median; Hinges: 1st& 3rdquartile; Whiskers: 1.5 x interquartile range; Points: outliers)

demonstrated a significant difference between the Intermediate and CCT groups (Appendices Table C.4).

Maximum acceleration

The diagnostic knee arthroscopy task demonstrated significant performance differences as measured by maximum acceleration at the all groups level ($p=0.00088$) and between the Intermediate and Novice group ($p=0.0011$), but not the CCT and Intermediate group.

Of the remaining tasks, only the Ring return task demonstrated a statistically significant difference at the all groups level ($p=0.00045$). As with the diagnostic knee arthroscopy, this was apparent between the Novice and Intermediate groups ($p=0.0014$), but not the CCT and Intermediate groups (Appendix C.5).

Stationary time, Inactivity time and Hand dominance

Neither the diagnostic knee arthroscopy exercise nor any of the other exercises demonstrated statistically significant differences at the all-groups level in the stationary time, inactivity time, or difference in hand dominance during the tasks (Appendix C.6, C.7, & C.8).

Survey responses

All participants completed the seven tasks in the simulated arthroscopy list. All of the Novice group (10; 100%), eleven of the Intermediate group (11/13; 85%) and nine of the CCT group (9/10; 90%), with 30 responses in total.

Metrics

The majority of participants felt the metrics to be motivating (25; 83%), with three (10%) participants ambivalent and two (7%) participants with no opinion toward them. None found the metrics to be demotivating.

In terms of perceived utility for training, the metrics which distinguished between participants based on experience above (Time, both Movement counts, and Smoothness) were generally more positively regarded (Table 4.5). Participants

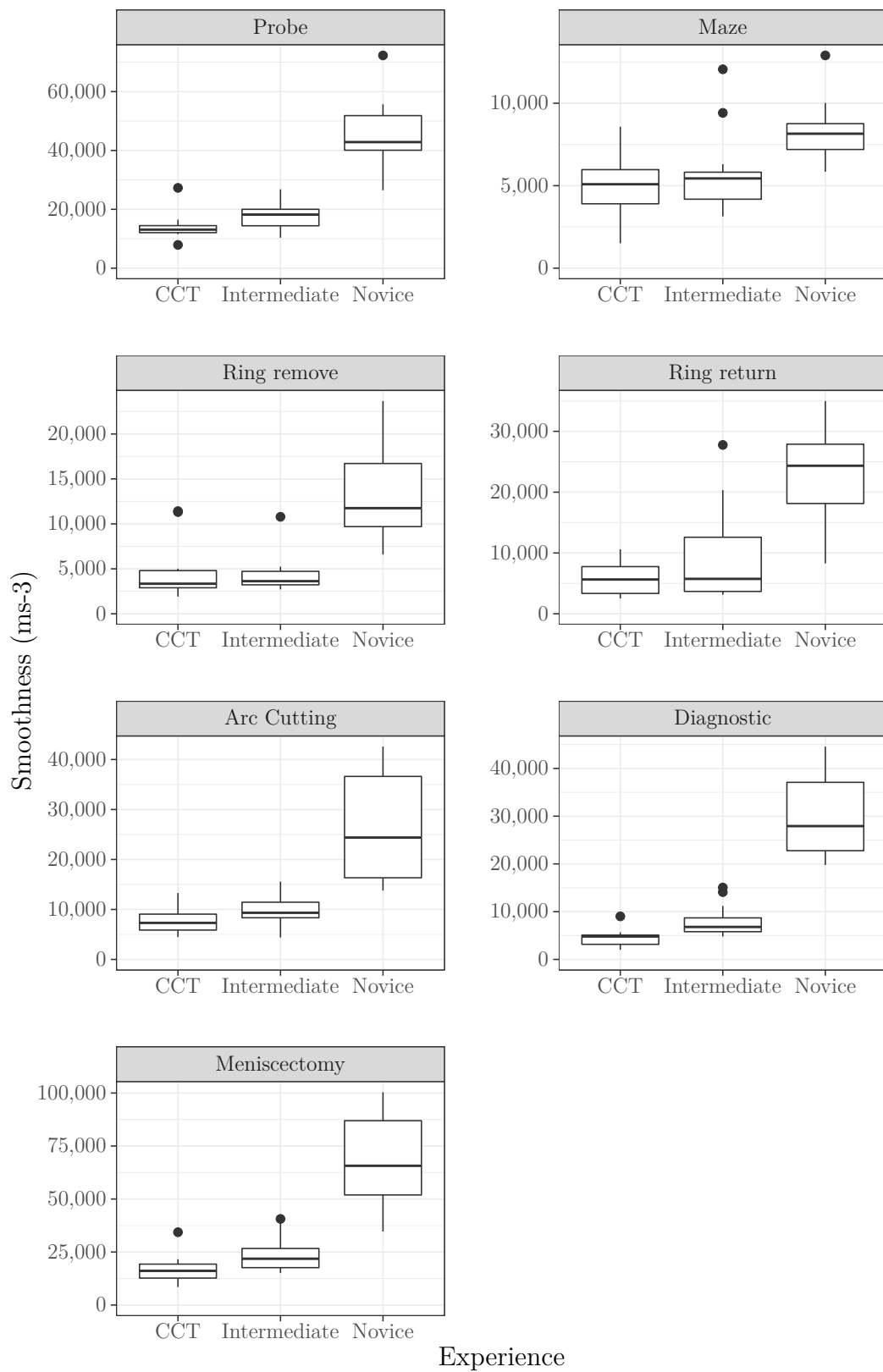


Figure 4.10: Performance on each simulated task measured by smoothness (ms^{-3}). (Centre bar: median; Hinges: 1st & 3rd quartile; Whiskers: 1.5 x interquartile range; Points: outliers)

rated the metrics more often as Useful or Potentially useful for Time taken (20/30; 67%), Movement counts (23/30; 77%), Smoothness (24/30; 80%) than for Maximum acceleration (16/30; 53%), Stationary/Inactivity time (18/30; 60%), and Difference in hand dominance (19/30; 63%).

Table 4.5: Usefulness: count of participant responses to the question “Having performed some arthroscopic tasks with motion analysis feedback, how useful do you anticipate these motion analysis parameters might be for training?”*

	Useful	Potentially useful	Neither	Potentially detrimental	Detrimental	N/A
Time	8	12	1	3	0	6
Movements	12	11	1	0	0	6
Smoothness	16	8	0	0	0	6
Max. Acceleration	6	10	7	1	0	6
Stationary/Inactivity	4	14	6	0	0	6
Hand dominance	8	11	5	0	0	6

* Full question stems and responses available in Appendix C.3

The metrics which distinguished between participants based on experience (Time, both Movement counts, and Smoothness) were also more frequently rated as Intuitive or Understandable (at least 83% of participants, compared to other metrics (57-70% of participants). They were also less frequently rated at Confusing or Irrelevant (0-13% vs 20-43%; Table 4.6).

Table 4.6: Clarity: count of participant responses to the question “Is it clear how the metrics represent performance?”*

	Intuitive	Understandable	Confusing	Irrelevant	No opinion
Time	18	7	3	1	1
Movements	15	14	1	0	0
Smoothness	13	17	0	0	0
Max. Acceleration	5	11	12	2	0
Stationary/Inactivity time	9	11	9	1	0
Hand dominance	8	13	6	2	1

* Full question stems and responses available in Appendix C.3

Most participants felt they knew how to improve their performance as measured by Time taken (20/30; 67%) and Movement counts (23/30; 77%), but fewer than half of participants felt they knew how they could improve performance on the other metrics (Table 4.7).

Table 4.7: Improvement: count of participant responses to the question “Is it clear how you could improve your performance?”*

	Know	Understand	Can't interpret	No opinion
Time	20	5	0	5
Movements	23	4	0	3
Smoothness	13	14	0	3
Max. Acceleration	7	8	10	5
Stationary/Inactivity time	12	13	2	3
Hand dominance	8	12	5	5

* Full question stems and responses available in Appendix C.3

FAST tasks

All participants felt that training on the FAST tasks would tend to improve triangulation, spatial perception and manual dexterity and most (23/30; 77%) felt it would tend to improve tissue manipulation. Most (16/30; 53%) felt it would have no, or minimal, impact on anatomical knowledge (Table 4.8).

Table 4.8: Impact of FAST: count of participant responses to the question “What impact do you think time spent on this FAST workstation would have on these general arthroscopic skills?”

	Tend to improve	None or minimal impact	Tend to impair
Anatomical knowledge	14	16	0
Triangulation	30	0	0
Spatial perception	30	0	0
Tissue manipulation	23	7	0
Manual dexterity	30	0	0

Meniscectomy task

When asked about the impact of the meniscectomy task on arthroscopic skills, at least 80% (24/30) participants felt it would tend to improve Triangulation, Spatial perception, Manual dexterity, Identification of intra-articular structures, Instrument handling, Inspection with the arthroscope, Navigation and probing, and Meniscectomy skills. At least a third of participants felt it would have no, or minimal, impact on Anatomical knowledge, Patient and equipment preparation, Portal placement, Insertion of the arthroscope through the portal, or on Meniscal

suturing. A small number felt it would tend to impair portal placement (2/30;7%) or insertion of the arthroscope through the portal (1/30; 3%) (Table 4.9).

Table 4.9: Impact of meniscectomy task: count of participant responses to the question “What impact do you think time spent on this meniscectomy simulator would have on these general/specific arthroscopic skills?”*

	Tend to improve	None or minimal impact	Tend to impair
Anatomical knowledge	21	9	0
Triangulation	30	0	0
Spatial perception	29	1	0
Tissue manipulation	27	3	0
Manual dexterity	30	0	0
Identification of structures	24	6	0
Instrument handling	30	0	0
Patient/equipment preparation	9	21	0
Portal placement	10	18	2
Inspection with arthroscope	26	4	0
Navigation and probing	30	0	0
Insertion through portal	16	11	1
Meniscectomy	27	2	0
Meniscal suture	19	11	0

* Full question stems and responses available in Appendix C.3

Regarding the realism of the model, the novice group were excluded due to insufficient experience. The arthroscopic display, arthroscope and instruments were felt by all to realistic (Table 4.10). Although participants felt some elements were unrealistic, none felt this detracted from the purpose of the simulation.

Table 4.10: Realism: count of participant responses to the question “What is your opinion of the realism of the following features of the meniscectomy simulation?”*

	Realistic	Unrealistic, not detracting	Unrealistic, detracts	Don't know/no opinion
Texture of structures	18	2	0	0
Colour of structures	14	6	0	0
Size of structures	17	3	0	0
Size of joint space	13	7	0	0
Arthroscopic display	20	0	0	0
Arthroscope	20	0	0	0
Instruments	20	0	0	0
Feel when probing	16	4	0	0
Fell when cutting	13	7	0	0

* Full question stems and responses available in Appendix C.3

All participants felt the meniscectomy model was Obvious or Implied to be a knee joint, and 87% (24/30) of participants felt all the essential anatomical features were present. Given the opportunity to elaborate on missing features, three participants responded (“fat pad, bleeding”, “capsule”, and “there is a lot more space in the simulator unlike in the real joint”).

Use in training

All participants, bar one in the CCT group, said they would have used the FAST tasks during their training if available (29/30; 97%). All participants, bar one in the Intermediate group, said they would have used the meniscectomy task during their training if available (29/30; 97%). When asked at which stage the SAL would be suitable for training, novice level trainees were most commonly suggested (29/30; 97%). Intermediate level trainees were suggested by as many from the Novice and Intermediate groups, but fewer of the CCT group (Table 4.11).

Table 4.11: Use in training: count of participant responses by experience group to the question “At what stages do you think this simulator is a suitable training modality in the skills lab?”

Group	Appropriate for Novices (e.g. Medical Students, Foundation and Core Trainees)	Appropriate for Intermediates (e.g Registrars)	Appropriate for Experts (e.g. Fellows and Consultants)
CCT	9/9	06/9	1/9
Intermediate	10/11	10/11	3/11
Novice	10/10	10/10	4/10

4.4 Discussion

4.4.1 Implications of the studies

Four FAST modules were identified as being appropriate for junior trainees, namely Probe, Maze, Flat Ring transfer, and Arc cutting tasks. Together with the diagnostic and meniscectomy knee arthroscopy tasks, these tasks made the ‘simulated arthroscopy list’ (SAL). The diagnostic knee arthroscopy task acted as a reference task, indicating that the time taken, number of hand movements, number of minor

hand movements, smoothness and maximum acceleration metrics could potentially differentiate participants based on experience in conjunction with the novel tasks. All of the tasks except the Maze demonstrated construct validity when performance was measured using the time taken, number of hand movements, number of minor hand movements and smoothness metrics. The SAL list was well received by all levels of participant, with the majority stating they would have used the FAST and meniscectomy tasks during their own training, and that the SAL overall would be appropriate to novices trainees. Participants felt that the time, movement counts, and smoothness metrics (which all demonstrated construct validity) were more useful metrics, clearer at representing performance, and they understood better how they could improve on their performance compare to the other metrics. Participants generally felt the FAST and meniscectomy tasks would improve performance of technical skills, but not anatomical knowledge, or patient preparation and portal placement: these tasks were not simulated and require a degree of decision-making which is beyond the scope of the FAST workstation. Recently, specific simulators have been designed to address the haptics behind the safe creation of arthroscopic portals, but not the decision-making behind portal placement and it's consequence for access during the rest of the case.²⁹⁵

When performance on the diagnostic knee arthroscopy task was assessed using the metrics produced by the wireless elbow-worn motion sensors, the time taken, number of hand movements, number of minor hand movements, and smoothness metrics were able to distinguish between Novice, Intermediate and CCT group, while the maximum acceleration metric distinguished between Novice and Intermediate groups. This indicates construct validity. The stationary time, inactivity time, or difference in hand dominance metrics did not detect any performance differences with experience, indicating lack of construct validity. When performance on the FAST and meniscectomy tasks was assessed, the time taken, number of hand movements, number of minor hand movements and smoothness metrics were consistently able to distinguish between Novice and Intermediate groups in all tasks except the Maze, indicating construct validity at this experience level, but this was not consistently

seen between the Intermediate and CCT groups. The maximum acceleration, stationary time, inactivity time, or difference in hand dominance metrics did not detect any performance differences with experience in the majority of the tasks, indicating lack of construct validity.

Based on these findings, it was decided that the simulated arthroscopy list for deliberate practice by novice trainees would consist of the Probe, Maze, Ring remove, Ring return and Arc cutting FAST tasks, with the diagnostic knee arthroscopy and meniscectomy anatomic tasks, combined with proximate performance feedback provided by the time taken, number of hand movements, number of minor hand movements and smoothness metrics. These metrics were those which most consistently distinguished participants based on their experience on the novel tasks and the validated diagnostic knee arthroscopy reference task. Although the Maze task did not demonstrate construct validity, it was rated as top three most useful exercises by all participants in study A and is the only task designed to specifically practice ‘tracking and probing a moving target’, a core arthroscopic competency.⁹⁸ Its inclusion is qualified by the recognition that participants may not be able to get meaningful feedback on their performance during deliberate practice when using the selected metrics, but is also a recognition that performance is not necessarily learning.¹⁰⁷

The main strength of these two studies is that they approach the tasks and metrics for the purpose of deliberate practice by a trainee. Although some studies do have multiple tasks which could permit deliberate practice,^{268,281,296} when assessing their construct validity the metrics are used to highlight the ability to ‘test’ participants skill in high-stakes assessments. The metrics presented here have been assessed for their ability to act as a guide for trainees, by asking how easy these metrics are to understand, how clearly they relate to performance, and how they can be used to guide improvement in training. When other studies have developed tasks and metrics expressly for the purpose of training, this has been to define performance thresholds for proficiency-based training, essentially a form of assessment, reached through simple repetition.^{129,291} Global rating scales, task-specific checklists and error lists rely on expert insights into high-quality performance,^{128,255} and one cannot

expect trainees without that experience to have the insight to be able to interpret why these factors indicate performance or how to achieve them. For example, Angelo et al used task-specific checklists and error lists to assess proficiency to progress to the final assessment in their studies, but participants received their direct proximate feedback on performance and suggestions on how to improve by faculty members.¹²⁹ Metrics cannot be effective for independent deliberate practice if they require external interpretation.

This study does not complete its own task deconstruction to identify the specific arthroscopic skills required for trainee surgeons and developing simulations to address these, but the FAST workstation was specifically designed to for this role,²⁶⁵ supported by the AANA and ABOS, and is widely-used in residency programmes in the United States. The simulated arthroscopy list has a variety of generic and anatomic tasks for deliberate practice by trainees because they are targeted to the aims of the curriculum for most orthopaedic surgeons. The simplest tasks that would achieve the goals of the curriculum were selected because inappropriate complexity of training is unnecessary and can be detrimental to performance and engagement in training.¹⁰⁵ Post-CCT fellows were asked to decide on the tasks because they have insights to the requirements and difficulties of arthroscopic training having just completed it themselves: long-standing consultants will have the required technical skills, but the environment in which they learned these skills is likely to be very different to that experienced by a modern trainee. As meniscectomy is part of the curriculum it is important to include it in the simulated arthroscopy list: a novel model was used in favour of a commercially available model as neither had yet been validated, and costs need to be controlled in simulation.

4.4.2 Limitations

The testing of multiple metrics for each of the selected tasks increases the risk of a Type I error (i.e. a false positive). The metrics assessed each had a reasonable hypothesis that they would be a valid measure of performance: time taken and counts of movements have been frequently validated elsewhere (Chapter 2); smoothness

(i.e. rate of change of acceleration, or jerkiness) and maximum acceleration could detect the nature of the movements counted and the potential for injury;^{232,297} the percentage of stationary time and difference in hand dominance may indicate bimanual coordination which is recognised as a measure of arthroscopic skill;^{203,249,255} as is inactivity time, which represents failure to plan and progress in the task.²⁹⁸ Despite reasonable belief they could be relevant measures of performance, assessing each metric against multiple tasks increases the family-wise error rate. To counter this, the accepted level of accepted significance for each metric was adjusted for each of the seven tasks by the Bonferroni method to reduce the Type I error rate.²⁹⁹ There is also a case to propose seven tasks, eight metrics, and all-group and both group-pair comparisons constitutes a separate hypothesis to correct for: this gives an effective level of α of 0.0003, which increases the chance of Type II error.

The decision against such a conservative correction is supported by the fact that statistically significant results are not randomly distributed throughout all metrics, tasks, and tests, but are well structured. Positive results were not found in any task for the stationary time, inactivity time and difference in hand dominance metrics, and only in two tasks for the maximum acceleration metrics. Positive results tended to be found for the time taken, number of hand movements, number of minor hand movements, and smoothness metrics between the Novice and Intermediate groups, except for the Maze task which was consistently unable to differentiate between participants by any metric. This structure where the findings largely consistent for each metrics, tasks and between group pairs suggests these are not false positive results, and that further correction for such correlated tests would be too conservative.²⁹⁴

4.4.3 Summary

These two studies support the use of a simulated arthroscopy list composed of the probe task, maze task, ring remove task, ring return task, arc cutting task, diagnostic knee arthroscopy task and meniscectomy task combined with the time taken, number of hand movements, number of minor hand movements and smoothness motion

analysis metrics for skills assessment and the deliberate practice of the arthroscopic skills required by early stage trainees. A validated list such as this might be integrated into early surgical training to help pre-patient training of fundamental arthroscopic skills before embarking on arthroscopic procedures in theatre. This training environment provides the opportunity to explore the arthroscopic learning curve and how it is affected by deliberate practice with objective feedback to make pre-patient training more efficient, and an opportunity to objectively investigate the transfer independent simulation training into intra-operative performance.

Practice doesn't make perfect. Only perfect practice makes perfect.

— Vince Lombardi

5

Randomised controlled trial to assess the impact of objective performance feedback on the learning curve during simulated arthroscopy practice

Contents

5.1	Introduction	129
5.1.1	Aim and hypothesis	131
5.2	Methods	132
5.2.1	Ethical approval	132
5.2.2	Participants and study setting	132
5.2.3	Simulation sessions, task, and feedback	132
5.2.4	Power calculation	134
5.2.5	LC-CUSUM	135
5.2.6	Outcomes and statistical analysis	136
5.3	Results	137
5.3.1	Baseline performance	137
5.3.2	Rate of learning: differences between groups each week .	139
5.3.3	Rate of learning: first attempt reaching 'consultant level' performance	139
5.3.4	LC-CUSUM	140
5.3.5	Performance change without knowledge of results	140
5.4	Discussion	143
5.4.1	Implications of the study	143
5.4.2	Limitations	144
5.4.3	Summary	147

5.1 Introduction

Learning curves describe how task performance improves with time³⁰⁰ and are recognised to exist when new surgical skills are acquired.^{34,301} They are a confounder in clinical trials of surgical innovations,^{27,30,167} and for trainees learning a surgical procedure they manifest as slower operating times^{28,29} and increased rates of complications,^{25,112,121} although they have not been linked to worse patient-reported outcome measures at long-term follow-up. Surgical training programmes in many countries mandate a number of supervised procedures before entering independent practice,^{302,303} but this is not tailored to the individual trainee's needs and there is wide variation of the estimated case-volume required to reach competence in various arthroscopic procedures^{125–127} and the ability to training programmes to provide enough opportunities.⁸³

One role of simulation training is to let trainees “learn the skills and prove their competence outside the clinical setting”⁹⁹ moving the early, hazardous stages of the learning curves out of the patient-facing environment.^{113,291} Performance metrics have been used to monitor the surgical learning curve and identify thresholds for competent performance, and simulation can be more efficient use of a trainee's time than theatre.⁷³ However, there is no surgical research on how to actively modify or optimise the learning curve in or out of theatre. Feedback is an essential part of skills learning,^{87,100} but the context of the feedback can positively or negatively affect learning rates and retention. Trainees can become dependent on ‘knowledge of results’ so performance may deteriorate when the feedback is removed³⁰⁴ when moving from the skills lab to the operating theatre, for example, and focusing on performance goals can prevent trainees from learning the processes to reach that goal.¹⁰⁵

Simulation can potentially be expensive,^{153,154} and without evidence to understand how simulation modifies the learning curve, affects intra-operative performance,

or modifies patient outcomes, one cannot assess its value for money. Although simulation may save money for the host institution where trainees slow workflow,^{27–29} it adds to the financial cost of those responsible for medical training: in the UK, this department is set to lose 20% funding by 2020,⁸² so there is a duty to ensure investment in simulation is done as efficiently as possible. Higher levels of investment or more senior input during simulation are not practical in the current environment.

Inferring arthroscopic learning curves from existing studies is difficult. Construct validity studies typically measure only single attempts at simulated tasks by practitioners at various training stages so we cannot see the learning curve of individuals, only the impact of their previous training. Where transfer validity has been reported in arthroscopy training, sequential performance with simulation training is not always recorded^{145,185} or the same methods of performance assessment are not used in the simulated and intra-operative settings^{144,145,185} so one cannot link the learning curve to final performance. Where studies do aim to investigate the learning curve itself (i.e. changing performance with sequential attempts) they have often had methodological or statistical flaws.

Methodologically, previous learning curve studies do not optimally represent a training environment that can allow deliberate practice. Learning curve studies have monitored skill acquisition in the ‘work’ environment by trainees²⁵⁸ or by qualified surgeons.^{126,179,196,206,208,209,222,225,226,245} These studies have used inappropriate metrics such as time, a process outcome which poorly represents surgical skill²³ and provides perverse incentives for training,^{164,305} or global rating scales which can’t be used independently to guide practice.²⁵⁸ Even when conducted in a simulated setting which might allow deliberate practice, the mode of practice is suboptimal, involving simple repetition without performance feedback,^{195,221,234,247} or instructional coached practice^{129,306–308} which is impractical if simulation is to become more wide-spread.^{153,154} As such, many studies are not suitable to describe the impact of deliberate practice with objective feedback on the learning curve.

Statistically, the approaches used to investigate the learning curve are often inappropriate. The chosen outcomes can be surrogates for performance, such a

complication rates which, as rare events, are insensitive to failures.^{23,121} Visual analysis of the outcomes such as plotting cumulative failures are descriptive and lack a formal test of when required performance is reached.¹¹¹ Moving-average performance unfairly penalises trainees biasing toward the contribution of past performance when we are interested in current performance.^{309,310} The CUSUM method, an industrial statistical process control method which has gained popularity in medicine for monitoring performance,^{311–313} defines the mean adequate and inadequate steady-state performance to calculate a test statistic to detect when performance deviates from an ‘in-control’ to ‘out-of-control’ state. CUSUM has been co-opted to investigate learning curves,¹¹¹ but because trainees performance is ‘out-of-control’ to begin with, it violates the assumptions of the test statistic.³⁰⁹ The LC-CUSUM is a variant of the CUSUM specifically addressing the violated assumptions to create a method to detect when trainees move from an ‘out-of-control’ to an ‘in-control’ state.³¹⁴

5.1.1 Aim and hypothesis

This study aims to investigate whether deliberate practice with automated feedback can be used to improve the learning curve in simulated skills acquisition. The hypothesis is that participants who undertake deliberate practice with automated feedback will acquire simulated skills more quickly than those who undertake unguided practice. By measuring skill using objective motion analysis metrics, participants are expected to show an improved rate of skills acquisition by reaching ‘consultant level’ performance and reaching an ‘in-control’ state earlier in the course of the study.

Secondarily, this study will also investigate whether performance is impaired when knowledge of results is removed. The hypothesis that the performance of participants undertaking deliberate practice with automated feedback as measured by the same metrics will deteriorate when automated feedback after each attempt is removed.

5.2 Methods

5.2.1 Ethical approval

The study was conducted under ethical approval from University of Oxford CUREC (MSD-IDREC-C1-2014-152; Appendix A.4, pg. 191).

5.2.2 Participants and study setting

Participants were medical students at the University of Oxford. Students with previous experience of arthroscopic or laparoscopic surgery or simulators were excluded. Participants gave informed consent to the study before viewing an orientation PowerPoint presentation. This presentation explained how to complete the task, that the purpose of the study was to see how objective feedback affected the learning curve, and that they should complete the task in an “efficient, accurate, and timely manner”.⁶³ Participants were randomised in a 1:1 ratio by computer to receive immediate performance feedback (see below) or no feedback between attempts. Randomisation took place after the first task attempt so as not to prejudice their baseline performance.

A randomised controlled trial took place at the OOSEC skills lab over six-month period. Simulated arthroscopy was performed in the skills laboratory using a standard 30° arthroscope, Smith & Nephew 460H camera, light source, and image management system, with images displayed on a 15-inch Iiyama TFT monitor. The equipment and simulation station arrangement is described in Appendix A.

5.2.3 Simulation sessions, task, and feedback

Each participant made four visits at weekly intervals to the OOSEC skills lab and completed 10 attempts at the task each time. In the first three weeks (attempts 1-30), participants received any performance feedback after each attempt according to their randomisation, to investigate whether performance diverged between the groups. In the final week (attempts 31-40) neither the group received feedback after any attempt, to investigate the effect of knowledge of results on the group who had previously received feedback. All arthroscopy was performed using a

standard 30° arthroscope, high definition camera and screen, while wearing wireless elbow-worn motion sensors. A supervisor (the author) was present throughout to operate the equipment, but not to provide coaching.

The Flat ring transfer FAST task was used, combining the Ring remove and Ring return tasks into a single event as originally developed by ABOS and AANA.³¹⁵ Each green ring is removed from the vertical metal pegs and placed in the centre well, then the green rings are returned to the vertical pegs. This task was chosen because it is similar to the FLS peg transfer task used by Feldman et al³¹⁶ and received the most ‘advanced’ ratings from post-CCT fellows during the task selection study (Chapter 4). It was hoped this would indicate a more challenging task and ‘prolong’ the learning curve to make intra-group differences easier to detect.

Sessions were set at 10 task attempts so each visit could be completed in approximately 30 to 60 minutes. The American Board of Orthopaedic Surgery Fundamental of Arthroscopic Surgery Training curriculum estimates ‘box training’ (generic bench-top simulation) should take up to two hours.³¹⁷ Drawing on work by Feldman et al³¹⁶ using a similar peg transfer task in medical students also, all participants recorded at least one ‘senior’ level score within 12 attempts, but CUSUM analysis of the same cohort showed only 1 participant (6%) demonstrated ‘senior’ level competency over 40 attempts, compared to 10 (63%) at ‘junior’ level.³¹⁸ Using LC-CUSUM in a knee arthroscopy task with more experienced participants, Hodgins et al found that the median number of attempts to competency was 16, albeit using a different task.²⁵⁸ Thirty attempts was chosen for parity with ABOS recommendations³¹⁷ and previous work^{316,318} to allow sufficient time for skills to develop without over-burdening volunteers.

After each attempt, participants in the non-feedback group proceeded on to the next attempt until 10 attempts had been completed for that session, whereas participants in the feedback group were shown the WASP client ‘dashboard’ of their performance, where they could review the number of hand movements, number of minor hand movements, smoothness and time taken to complete the task (Figure 5.1). They could view their performance relative to a ‘consultant level’ (the mean

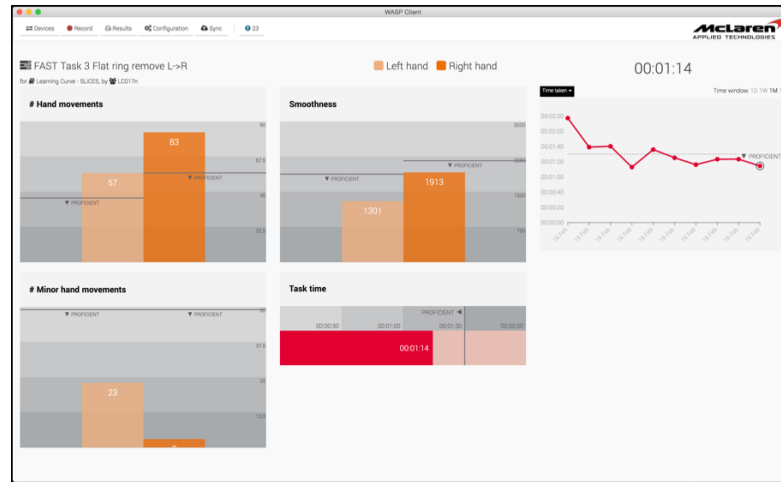


Figure 5.1: The WASP client ‘dashboard’ allowed participants to view results feedback immediately after each attempt and view progress over time.

performance of the CCT group in Chapter 4, Study B) and compare their current performance to previous attempts. Participants in the feedback group did not receive guidance on how to interpret the dashboard. No participants received coaching on how to improve their performance, but were reminded at each attempt to complete the task in an “efficient, accurate, and timely” manner. If participants asked for assistance, they were reminded that the purpose of the study was to simulate independent practice. If they voiced their own interpretations of the metrics or how they might improve them, these were neither confirmed or refuted by the supervisor.

5.2.4 Power calculation

Using the construct validity study in Chapter 4 as a pilot, the results of the Ring remove and Ring return tasks were summed to get a total for number of hand movements, number of minor hand movements, smoothness, and time taken for the Flat ring transfer task to act as reference performance levels. It was anticipated that participant progress would improve with practice, so power was calculated based on the number of hand movements required by the CCT group to complete the task (mean 225; SD 50). A sample size of 34 was required to give an 80% chance (Type II error; $\beta = 0.2$) to detect a 1 SD difference (50 moves) in the mean number of hand movements to complete the task (two-tailed Type I error;

$\alpha = 0.05$). A 1 SD shift is equivalent to 22.2%, which is of the order of the 19% difference noted by Feldman et al.³¹⁶

5.2.5 LC-CUSUM

Learning curve cumulative sum (LC-CUSUM), a variant of the CUSUM statistic, is used to detect when a participant's performance moves from an 'out-of-control' (performing inadequately) to an 'in-control' state. It uses a defined adequate level of performance (μ_0), an inadequate level of performance (μ_1), and acceptable deviation from adequate performance (δ). The null hypothesis is that participants are 'out-of-control' which correspond to an average deviation from acceptable performance μ_0 by at least δ .

For each sequential attempt, the test statistic (S_t) is calculated where successful attempts decrease the value of S_t , while unsuccessful attempts increase S_t . S_t is set to 0 at the outset, and there is a holding barrier preventing S_t rising above 0 so the test remains responsive to improvements in performance and participants are not 'punished' for earlier accumulated failures. The null hypothesis that performance deviates from μ_0 by at least δ (i.e. more than an acceptable level of deviation) is not rejected while S_t remains above the predetermined threshold h . After each attempt the updated statistic S_t is compared to a threshold h and whilst S_t remains above h , the null hypothesis is not rejected and the participant is considered 'out-of-control'. When S_t passes the threshold h , the null hypothesis is rejected, and the alternative hypothesis that the participant is 'in-control' is accepted. From this point, the test statistic S_t is no longer informative about performance.

The parameters for LC-CUSUM are arbitrarily defined based on a consensus of what is acceptable. There is a lack of data on acceptable performance on the FAST workstation, and none beyond this thesis using wireless motion analysis metrics. In this study, the adequate level of performance (μ_0) was set to 'consultant level' performance (the mean performance of the CCT group in Chapter 4, Study B, which was being used as the reference performance level for feedback in the intervention group) and the proposed inadequate performance (μ_1) set to $\mu_0 + 20\%$,

Table 5.1: Selected h thresholds for each metric, and their respective probability of successful detection and probability of false alarm from 1000 simulations.

	h	Probability of successful detection at acceptable performance (alarm rate at μ_0)	Probability of false alarm at unacceptable performance (alarm rate at $\mu_0 + 20\%$)
Hand movements	2.60	0.90	0.05
Minor hand movements	2.90	0.89	0.05
Smoothness	N/A	N/A	N/A
Time	3.00	0.84	0.05

giving an acceptable deviation (δ) of 10%.³¹⁴ Due to the holding barrier at 0, the null hypothesis can never be accepted and as the number of attempts tends towards infinity, the false positive rate tends towards 100%. Because of this, the test is described using the probability of false alarm (PFA; analogous to false positive rate) and probability of successful detection (PSD; analogous to true positive rate) over a number of attempts. The threshold h was pre-determined for each metric based on 1000 numerical simulations of adequate and inadequate performance with the aim of giving a PFA under 5% and PSD over 80% after 30 attempts. The chosen threshold h for each metric are shown in Table 5.1 (see Appendix D for simulation results and effect of varying h on PSD and PFA). No h threshold was defined for the Smoothness metric, as the simulations showed no threshold h would permit PFA under 5% and PSD over 80%.

5.2.6 Outcomes and statistical analysis

The following performance metrics were recorded from the wireless elbow-worn motion sensors: number of hand movements, number of minor hand movements, smoothness, and time taken. For each metric, the outcomes of interest were differences in the learning rate for each group, and the changes in performance after removing knowledge of results (performance feedback) from the intervention group.

Learning rate was assessed by comparing performance at the end of each weekly session (every 10 attempts) and against two predefined standards: the attempt that participants first reached ‘consultant level’ performance, and the attempt that participants reached the h threshold (‘in-control’) using the LC-CUSUM statistic. Distributions of performance metrics at the end of each weekly session and the

attempt which they reached the two standards were assessed for normality using the Shapiro-Wilk test, and compared using the t-test or Mann-Whitney U test accordingly. The proportion of participants reaching the pre-defined standards was compared using the χ^2 test.

The effect of knowledge of results was assessed by removing performance feedback from the intervention group in the final weekly session (attempts 31-40) and longitudinally comparing performance on a per-participant basis at the end of the third session (attempt 30) to the end of the final session (attempt 40). Distributions was assessed for normality using the Shapiro-Wilk test, and compared using paired t-tests or Wilcoxon signed-ranks test accordingly.

A corrected p-value of 0.0125 (Bonferroni correction = $0.05/4$) was considered statistically significant to correct for the multiple comparisons (four motion metrics) being assessed. All statistical analyses were performed in R v3.3.2.²⁸⁴

5.3 Results

Thirty-four participants were recruited, with 18 randomised to the feedback group and 16 to the non-feedback group. Four participants had a client error during an attempt (Participant 005n, 010f, 011f and 024f during attempts 15, 26, 3 and 5, respectively), representing an equipment failure rate of 0.15% over the course of the study. One participant was in the non-feedback group, and three in the feedback group. Attempts where the client failed to record metrics were removed for analysis because the LC-CUSUM calculations cannot handle gaps in a series, so these participants appear to have had 39 attempts at the task. Mean participant performance for each metric is shown, with reference to mean ‘consultant level’ performance from pilot data in Chapter 4, in Figure 5.2.

5.3.1 Baseline performance

At the beginning of the first week, the two groups were well matched at baseline, with no statistically significant differences between the feedback and non-feedback groups for the number of hand movements ($p=0.34$), number of minor hand

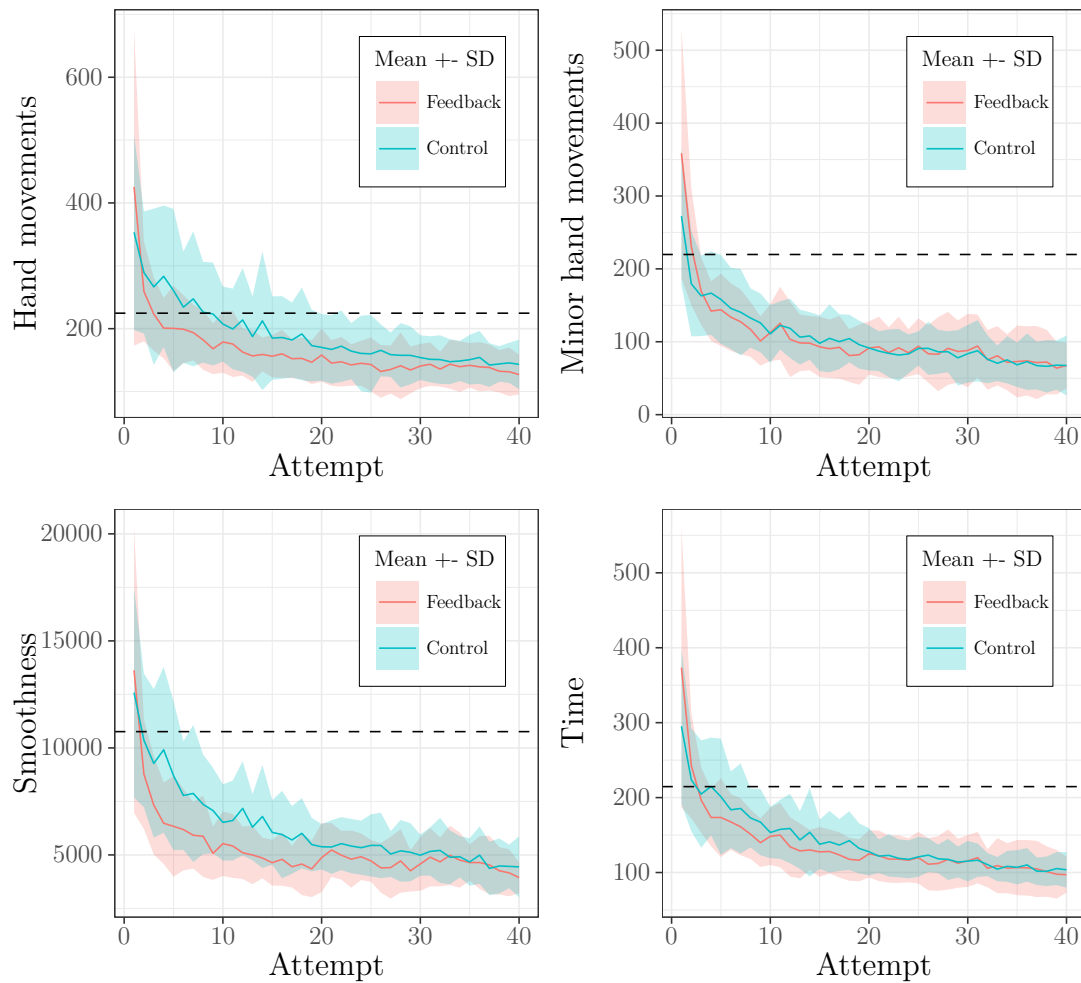


Figure 5.2: Mean participant performance (+/- standard deviation) at each successive attempt over the course of the study. Dashed line represents mean ‘consultant level’ performance for that metric at their first attempt at the task.

Table 5.2: Difference in performance between the groups at baseline.

Metric	Feedback group	Non-feedback group	
Hand movements (n)	353 (274 – 439)	309 (254 – 415)	p=0.34
Minor movements (n)	317.5 (257.5 – 399.5)	243.5 (204 – 331.5)	p=0.10
Smoothness (ms^{-3})	11671 (9419 – 14958)	11326 (8791 – 16475)	p=0.85
Time (s)	339 (269 – 403)	253 (229 – 375)	p=0.18

movements ($p=0.10$), smoothness ($p=0.85$), or time taken ($p=0.18$) to complete the task (Table 5.2).

Table 5.3: Difference in performance between groups at end of weeks 1-3.

Week	Metric	Feedback group	Non-feedback group	
1	Hand movements (n)	178 (146.75 – 202)	209 (184.75 – 227.5)	p=0.07
	Minor movements (n)	112.5 (39.0)	110.9 (39.5)	p=0.91
	Smoothness (ms ⁻³)	5525.7 (1508.4)	6519.9 (1783.7)	p=0.09
	Time (s)	148.4 (38.3)	153.4 (34.2)	p=0.70
2	Hand movements (n)	158.0 (42.2)	170.2 (53.0)	p=0.48
	Minor movements (n)	91.4 (36.6)	91.5 (30.9)	p=0.97
	Smoothness (ms ⁻³)	4872.0 (1610.0)	5381.6 (1060.1)	p=0.29
	Time (s)	124.9 (31.5)	127.7 (31.5)	p=0.81
3	Hand movements (n)	140.5 (35.2)	154.0 (35.5)	p=0.22
	Minor movements (n)	69 (55.5 – 109.5)	77 (58.5 – 103.5)	p=1
	Smoothness (ms ⁻³)	4581.1 (1253.1)	4987.8 (958.7)	p=0.37
	Time (s)	115.4 (34.6)	115.2 (27.0)	p=0.98

Results shown as median ($Q_1 - Q_3$), or mean (SD) if normally distributed

5.3.2 Rate of learning: differences between groups each week

At the end of the first, second and third week (attempts 10, 20, and 30, respectively), where the feedback group had received immediate feedback on their performance after each attempt, there was no significant difference between the groups as measured by the number of hand movements, number of minor hand movements, smoothness or time taken ($p > 0.0125$ in all cases; Table 5.3)

5.3.3 Rate of learning: first attempt reaching ‘consultant level’ performance

The serial performance of the participants for each metric relative to ‘consultant level’ performance are shown in Figure 5.2. All participants in both groups reached ‘consultant level’ performance by the end of the study so the proportions from each group were not assessed statistically.

The non-feedback group reached mean consultant performance for the number of minor hand movements after a median 1.5 attempts, earlier than the feedback group who took at median 3 attempts ($p = 0.0018$). There was no statistically significant difference in the attempt that the two groups reached ‘consultant level’ performance in the other metrics ($p > 0.0125$ for each metric).

Table 5.4: Comparison of the attempt number which participants from each group reached the h threshold to become ‘in-control’.

Metric	Attempt in-control; median ($Q_1 - Q_3$)		
	Feedback group	Non-feedback group	
Hand movements (n)	8 (5 – 11)	10 (7 – 16.5)	p=0.18
Minor movements (n)	4.5 (4 – 6)	4 (3.5 – 7.5)	p=0.50
Time (s)	6 (5 – 9)	8 (6 – 10)	p=0.27

5.3.4 LC-CUSUM

The proportion of participants reaching the h threshold over the course of 30 attempts was similar between the groups. One participant in the control group failed to become ‘in-control’ for the number of hand movements, but this was not statistically significant (p=0.95). All participants passed the h threshold for the number of minor moves and time taken within 30 attempts (Figure 5.3).

The attempt number which participants reached the h threshold was not statistically significantly different between the intervention and control groups (p>0.0125 for each metric; Table 5.4)

5.3.5 Performance change without knowledge of results

To assess the impact of knowledge of results, in the fourth week during attempts 31-40, the intervention group did not receive feedback between these attempts. Between attempt 30 (last attempt in which they had received feedback) and their final attempt, the feedback group showed no difference in the number of movements (p=0.53), number of minor moves (p=0.016), or smoothness (p=0.20). There was an average improvement of 12 seconds in the time taken (mean of difference = 12.0, 95% CI = 4.5, 19.5; p=0.0040) between attempt 30 and the final attempt (Table 5.5).

The control group continued to practice without feedback between attempt 30 and their final attempt, showing no statistically significant improvements in the number of hand movements, number of minor hand movements, smoothness or time taken (Table 5.5).

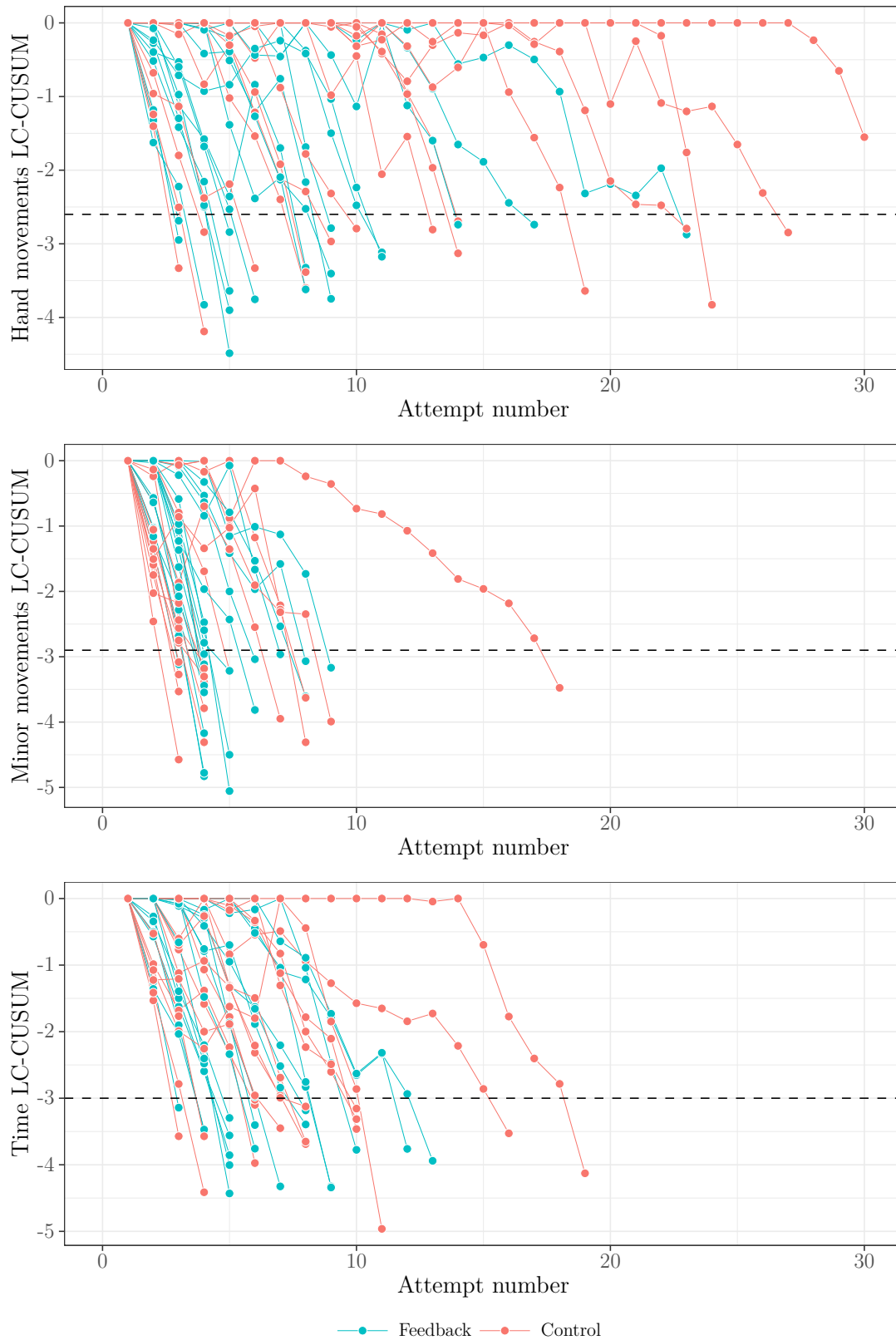


Figure 5.3: The connected scatter plots for each metric show each participant's score after each task attempt. LC-CUSUM score starts at 0 and participant performance is assumed to be 'out-of-control' while the score is above the calculated h threshold (the dashed line; also see Table 5.1). When a task attempt is within acceptable performance limits, the score improves (becomes more negative). A participant's performance is 'in-control' on the attempt where their score crosses the h threshold.

Table 5.5: Change in performance within groups in attempts 31–40 after the removal of feedback from the intervention group

Metric	Feedback group		Non-feedback group	
	at attempt 30	at final attempt	at attempt 30	at final attempt
Moves (n)	140.5 (35.2)	136.1 (38.3)	156.8 (35.1)	142.9 (37.9)
Minor moves (n)	69.0 (55.5 – 109.5)	66.0 (37.0 – 87.5)	75 (54 – 112)	67 (36 – 88)
Smoothness (ms ⁻³)	4581.1 (1253.0)	4242.1 (965.1)	4960.899 (989.0)	4415.2 (1376.3)
Time taken	115.4 (34.6)	103.4 (32.0)	115.6 (28.0)	103.6 (22.8)

Results shown as median ($Q_1 - Q_3$), or mean (s.d.) if normally distributed

p=0.029
p=0.079
p=0.086
p=0.078

5.4 Discussion

5.4.1 Implications of the study

This study was designed to assess the effect of unguided objective performance feedback on the learning curve, and the effect of removing that feedback on skills retention for subsequent attempts. In the first three weeks, differences in the learning curve were assessed by comparing absolute performance at defined time points and by comparing the attempt number when performance reached one of two thresholds (the first attempt to reach ‘consultant level’, and the h threshold using LC-CUSUM). There was no difference in absolute performance between the groups at the end of weeks 1 - 3 according to any of the objective metrics recorded. The control group reached ‘consultant level’ performance earlier than the intervention group according to one metric (number of minor hand movements; median 1.5 vs 3 attempts), but there was no difference between groups for the number of hand movements, smoothness or time taken metrics. Similarly, according to the LC-CUSUM score there was no statistically significant difference between the groups in the attempt number at which they reached the h threshold to become ‘in-control’ for any of the three metrics assessed (number of hand movements, number of minor hand movements, or time taken). In the final week, the effect of knowledge of results was assessed by removing feedback from the intervention group. There was no deterioration in performance over the course of the final 10 attempts, and time taken actually improved by an average of 12 seconds (10.4%) in the intervention group. The control group showed no statistically significant performance changes over the final week.

The learning curve is important to simulation training in the context of ‘pre-patient’ and proficiency/competency-based training, where deciding the point at which a trainee should progress to the next stage should be robust in terms of the metrics and statistics used.^{103,116,319,320} Ritter and Scott³²¹ used time and frequency of errors as metrics to define proficiency levels in the Fundamentals of Laparoscopic Surgery curriculum which are both poor measures of performance, but easy to

use during independent practice. Jackson et al²⁴⁷ use a global rating scales to monitor the learning curve and skills retention in arthroscopic meniscal repair, but these are inherently subjective and cannot be used independently. This study uses objective metrics that are meaningful measures of intra-operative surgical activity and could be calculated during independent practice. Using metrics to decide when trainees are proficient would typically involve participants demonstrating a single performance at a predefined ‘consultant’ level,^{129,198} an arbitrary number of repetitions at that level^{109,321} or a statistical passing score based on contrasting group performances.^{91,320} This study used the LC-CUSUM statistic, a method that is sensitive to the present performance of the trainee relative to the performance level required, and formally statistically assesses that performance is within acceptable limits. This study is unique in arthroscopic training by using LC-CUSUM in a ‘deliberate practice’ setting: other studies have monitored the learning curve during practice on patients while acquiring a new skill in a ‘work’ environment.^{226,258} The stresses and cognitive load of performing on real patients⁵⁰ and the inability systematically explore weaknesses make ‘work’ an inappropriate setting to explore the learning curve. This study aimed to optimise the metrics, statistics and setting to not only observe the learning curve, but understand whether it could be improved.

It is a recognised effect in motor skills training that practice with knowledge of results has been linked to improved skills acquisition, but impaired skills retention when the feedback is removed. It has been suggested that for more complex motor skills this effect does not apply,³²² but it’s role in surgical simulation has not been previously investigated. This study suggests that there is no impairment to practice performance following the removal of performance feedback, in keeping with findings for other complex motor skills.

5.4.2 Limitations

This study had some notable limitations, principally related to some of the disadvantages associated with the LC-CUSUM statistic and the rapid skills acquisition of the participants.

LC-CUSUM is designed to detect the point at which an individual participant's performance varies from adequate performance by no more than the acceptable amount of deviation: the alternative hypothesis (that performance is 'in control') is accepted, at which point it loses its statistical properties. Statistical comparison of the passing index (the attempt which one accepts participant performance is 'in control') is possible between groups, but this not the design of the statistic and LC-CUSUM does not have a specific group-level variant. One has to consider: are differences in passing index between participants statistically significant given LC-CUSUM is designed to monitor performance 'within' a given participant, and are the differences clinically relevant?

LC-CUSUM does not assess the depth (plateau) of learning beyond h .³⁰⁹ This means that, as in this study, where performance improves beyond the pre-defined acceptable level the LC-CUSUM does not provide any additional information or allow comparison between individuals, although the planned direct comparison of absolute performance between groups mitigates this. The h threshold is a decision limit of consistent performance relative to a standard, not a measure of performance in itself: adjusting the h threshold does not affect the level participants must reach, but influences the probability of successful detection and false alarms over the course of a defined number of attempts. Despite its statistical robustness, the LC-CUSUM statistic is not easily calculated, requiring pilot data on the group and task which might not always be available. Finally, it's presentation as a score is not intuitive: this study did not assess the use of S_t as feedback, but the negative scoring may be demotivating, undermining its utility as a performance metric for self-directed learning.

LC-CUSUM requires the adequate performance and acceptable deviation to be defined in advance. Where the data dispersion in the adequate performance group is large relative to the acceptable deviation, as was the case with the smoothness metric in this study, this affects the PSD and PFA. Given the available parameters, h was not defined for smoothness in this study as it was felt there was no acceptable trade off of PSD and PFA that would allow a useful assessment for the smoothness metric.

Using ‘consultant level’ data from Chapter 4 to define adequate performance was expected to be a challenging target for the participants with the aim to avoid a ceiling effect where participants passed too early and prevented any differences between the two groups becoming apparent. Participants in both groups acquired the skills more rapidly than anticipated and most surpassed consultant performance in all four metrics in the first weekly session. Because LC-CUSUM only detects when participants are ‘in control’ this in turn creates a ceiling effect and may contribute to the lack of differentiation between the study groups. Thus, although LC-CUSUM is a more statistically valid method to define when competence has been reached, in retrospect the performance metrics used may have impaired its effectiveness. *Post hoc* analysis could be made with an altered h threshold, but as discussed above this would alter the probabilities of false alarm and successful detection without changing the acceptable performance level. Using an alternative definition of adequate performance (e.g. the median performance of the participants at a certain point in the study) is both arbitrary and circular.

Only one task was chosen to be assessed. The simulated arthroscopy list was designed to allow the systematic exploration of weaknesses and deliberate practice of the sub-skills within basic arthroscopy.⁹⁸ The purpose of this study was not to see whether self-directed feedback improves the rate of learning on the simulated arthroscopy list, rather, the general concept of using metrics to improve a learning curve. The task choice was practical: it matches a common FLS task studied elsewhere;^{316,318} it involves both camera and instrument use; and is a more ‘advanced’ task of those available (See Chapter 4). Multiple tasks would make the study prohibitively long and the FAST tasks are sufficiently different to prevent direct comparison of performance, but have generalizable components so each task better prepares you for the next, which would complicate analysis.

Construct validity studies generally compare performance during a single attempt at a task. Intra-operative experience and simulator performance often correlate at the first attempt, but if simulated performance by novices can surpass initial expert performance after a few repeated attempts, this highlights the “distinction between

performance and learning”¹⁰⁷ and it undermines the notion that intra-operative experience and simulated practice have a commutative effect on learning and performance. This has implications for using simulation for high stakes assessment or credentialing if simulation training rapidly improves simulated performance without improving intra-operative performance in a predictable fashion. That is not to say there would be no improvement in intra-operative performance, just that after repeated practice the simulator may not be appropriate to assess progress. These issues cannot be addressed without objective assessment of the transfer validity of simulation training into the operating theatre.

5.4.3 Summary

This study shows that for this task, there is a marked impact of deliberate practice on performance, but self-directed objective feedback with wireless motion analysis did not improve the rate or depth of the learning curve over blind repetition. Importantly, the earlier presence of feedback was not detrimental to performance when feedback was unavailable.

The interaction between intra-operative experience, initial simulator performance, and repeated simulated performance is complex, and attempts to alter the shape of the learning curve are not the same as shifting the learning curve out of the operative setting: that requires an assessment of transfer validity to demonstrate that simulated activity affects intra-operative activity.

I know kung fu.

— Thomas A. Anderson a.k.a. Neo

6

Randomised controlled trial to objectively assess the transfer of surgical simulation training to intra-operative performance

Contents

6.1	Introduction	149
6.1.1	Aim and hypothesis	150
6.2	Methods	150
6.2.1	Ethical approval and registration	150
6.2.2	Participants and study setting	151
6.2.3	Equipment	151
6.2.4	Randomisation	152
6.2.5	Assessments and training	152
6.2.6	Outcome measures and statistical analysis	153
6.3	Results	155
6.3.1	Participant baseline simulated arthroscopy list performance	155
6.3.2	Participant in-theatre performance	158
6.3.3	Supervisor in-theatre performance	158
6.3.4	Participant:supervisor in-theatre performance ratio	159
6.3.5	Participant follow-up simulated arthroscopy list performance	163
6.3.6	Participant training activities during study	163
6.4	Discussion	163
6.4.1	Implications of the study	163
6.4.2	Limitations	167
6.4.3	Summary	169

6.1 Introduction

While technical ability is only one attribute of a good clinician, it is particularly associated with surgery:²¹ it affects patient outcomes^{25,27,33} so learning these skills is an essential domain of surgical competency.^{16,165,166} During surgical training, learning curves are expected when acquiring technical skills by trainees,¹¹⁸ trainers,¹²⁴ healthcare providers,³⁴ and patients,¹³³ yet the earlier stages of learning are linked with increased complication rates.^{25,112} Reduced working hours^{323,324} have coincided with modernised training programmes,^{53,79} resulting in reduced surgical exposure and time to train. The resulting extension of the learning curve beyond formal training can produce surgeons of tomorrow that are less experienced than their predecessors.³²⁵ There has been a shift from a time-in-training to a competency-based model in training programmes,^{22,78} but despite these changes trainee progression remains almost universally based on subjective assessment. Objective measures of technical skill to assess progress through training or to allow comparisons to competency benchmarks do not exist.^{11,87,118}

Surgical simulation is proposed on ethical grounds to move the learning curve out of the operating theatre,¹¹⁶ to allow more efficient skills practice^{73,99} and to mitigate working hours restrictions.^{79,152,324} Simulation is prominent in industries such as aerospace because objective measures of technical performance permit the calculation of training effectiveness ratios, where simulation costs correlate by a reproducible ratio to savings in flying time, fuel spent, and *in extremis* loss of aircraft and pilots.¹³⁵ In contrast, benefits from surgical simulation are more difficult to measure as surgical outcomes are not always proximate to the surgery itself or attributable to any single member of the surgical team.¹¹¹ Although some studies have implied the transfer of simulation training into improvements in intra-operative performance using subjective assessments like global rating scales,^{144,145,185,186} to date there is no objective evidence demonstrating clinical and cost-effectiveness of surgical simulation training.^{153,154}

Simulation and competency-based training need direct assessment measures of technical skill,¹⁰⁸ but objectively assessing the technical skills of surgeons and

surgical trainees performing procedures hasn't previously been feasible.^{11,87} Surrogate markers (e.g. complications or blood-loss) are typically rare or multifactorial, and are inappropriate proxies.²³ Psychometric assessments¹⁶⁰ are subjective and time-consuming for surgical trainers¹⁵⁴ who are under increasing pressure to deliver service and lower waiting times rather than teach and train future surgeons.³²⁶ As discussed in Chapter 2, many methods of motion analysis^{220,223,242,276–279} producing objective metrics used to measure technical performance have been limited to simulations by technical constraints. The wireless elbow-worn motion sensors validated in Chapters 3 and 4 can objectively assess surgical dexterity on simulators and in the operating theatre. When combined with the simulated arthroscopy list targeting basic arthroscopic skills these offer a unique opportunity to investigate the transfer of training.

6.1.1 Aim and hypothesis

The aim of this randomized controlled trial was to investigate whether surgical simulation training objectively improves the arthroscopic performance of surgical trainees in the real operating theatre. The hypothesis is that trainees who undertake surgical simulation training in addition to their usual deanery-provided training will acquire simulated surgical skills which will in turn transfer to better performance in a real operating theatre compared to those who undertake their usual deanery-provided training alone. By measuring simulated and intra-operative skill using objective motion analysis metrics, trainees are expected to show simulated skill acquisition and transfer of training into the operating theatre by greater improvement on the simulated practice tasks over the course of the study and better intra-operative performance at the end of the study.

6.2 Methods

6.2.1 Ethical approval and registration

The study was conducted under ethical approval from University of Oxford CUREC (MSD-IDREC-C1-2014-152; Appendix A.4, pg. 191). Participants gave informed

consent, and could withdraw at any point without detriment to their training. The trial was registered with the *Biomed Central ISRCTN Registry* (ISRCTN27154417) and *clinicaltrials.gov* (NCT02777333).

6.2.2 Participants and study setting

We conducted a randomised controlled trial from January 2016 to March 2017 according to CONSORT guidelines to assess whether simulation training during surgical training improves surgical technical performance in the operating theatre. Senior house officer trainees with GMC-approved Trauma & Orthopaedic training rotations at a UK teaching hospital were eligible for inclusion. Trainees with more than 2 years of surgical training; previous admission to a higher surgical training programme; or who had performed or assisted in over 10 arthroscopic/minimal-access procedures were excluded. Study performance was distinct from their employment, and not assessed by the employer or Deanery.

Simulated arthroscopy was performed in the skills laboratory using a standard 30° arthroscope, Smith & Nephew 460H camera, light source, and image management system, with images displayed on a 15-inch Iiyama TFT monitor. The equipment and simulation station arrangement is described in Appendix A. All operative procedures took place in the operating theatres of the Nuffield Orthopaedic Centre, a university teaching hospital.

6.2.3 Equipment

Elbow-worn wireless WAX9 motion sensors streamed three-axis accelerometer and gyroscopic data at 50Hz via Bluetooth to a laptop running the purpose-built client (McLaren Applied Technologies, Woking, UK). The client used these data to produce motion metrics of surgical hand dexterity, including the number of hand movements (n), number of minor hand movements (n), smoothness (ms^{-3}), and time taken (s). Simulation sessions took place on a diagnostic knee arthroscopy (DKA) model (#1401; SawBones, Pacific Research Laboratories, WA), and on six additional simulated tasks selected from the commercially-available FAST workstation developed

by the American Board of Orthopaedic Surgery and Arthroscopy Association of North America²⁹³ and validated as described in Chapter 4.

6.2.4 Randomisation

After obtaining written informed consent and a baseline simulator assessment, participants were randomly allocated (1:1) by computer to receive only their normal deanery-provided training (control group), or additional simulation training (intervention group). Participants could not be blinded to their allocation.

6.2.5 Assessments and training

The basics of an arthroscopic stack and 30° arthroscope, the instruments and simulators used, the relevant knee anatomy, and how to perform each of the tasks were demonstrated to all trainees during an orientation session at the beginning of the study. Participants then completed a baseline simulator assessment on the validated simulated arthroscopy list before randomisation so their allocation did not bias baseline performance. After 13 weeks of standard training (control group), or with additional simulation training (intervention group), participants completed a supervised DKA in the operating theatre supervised by a consultant knee surgeon. Participants then performed a final follow-up simulator assessment.

Standard deanery-provided training is managed for each trainee by two senior clinicians; a clinical supervisor and an academic educational supervisor. Duties and educational opportunities include day-time and out-of-hours ward management of surgical patients, attending outpatient clinics and operating theatres, teaching by seniors, and protected time for a regional teaching programme. The intervention group completed an hour of simulation training each week for 13 weeks on top of their standard deanery-provided training *ad hoc* around clinical duties. The training was self-directed, and participants used motion analysis to see how they performed on each task relative to their peers and seniors, allowing deliberate practice on specific arthroscopic skills.

For the supervised in-theatre assessment, both the participant and supervising consultant scrubbed wearing elbow-worn wireless motion sensors. As usual during standard training, the supervisor demonstrated a standardised DKA using an approved and previously validated sequence^{84,144}(Figure 6.1). The participant then repeated the same DKA. The supervisors were blinded to participants' randomisation. Criteria for selecting appropriate operative cases were agreed by faculty (Appendix E.1, pg. 226).

6.2.6 Outcome measures and statistical analysis

The wireless motion analysis system reported four objective metrics during the simulated arthroscopy list in the skill lab and during the diagnostic knee arthroscopy in the operating theatre: number of hand movements, number of minor hand movements, smoothness of movements, and time taken. Additionally, in theatre, a 'performance ratio' is calculated where each metric is expressed as a ratio of the participant's performance to the supervisor's performance while demonstrating the diagnostic round during the same case. This 'performance ratio' accounts for patient factors which could affect case complexity and performance, by presenting participant performance relative to the presumed 'optimal' performance of the senior surgeon performing the same case.

The primary outcome measure was the number of hand movements required to complete a diagnostic knee arthroscopy in theatre. The secondary outcomes were:

- The number of minor hand movements, smoothness and time taken to complete a diagnostic knee arthroscopy in theatre
- The 'performance ratio' of each metric during a diagnostic knee arthroscopy in theatre
- The number of hand movements and minor hand movements, smoothness and time taken during the simulated arthroscopy list tasks at follow-up

Insertion

1. Scope inserted through lateral portal cannula into supra-patellar pouch with knee in extension

Setting up

2. Orientation period: camera position, focus, and general 3D orientation

Inspection phase (*structure visualised to be deliberate focus of view*)

3. Identify trochlea
4. Identify patella. Visualise all facets
5. Swing back down medial gutter to enter medial compartment while flexing knee (supervising clinician to assist with positioning on hip)
6. Identify medial femoral condyle
7. Identify medial tibial plateau
8. Identify medial meniscus (posterior horn)
9. Enter intercondylar notch and identify anterior cruciate ligament
10. Enter lateral compartment following intercondylar notch (supervising clinician to assist with positioning in figure-of-four)
11. Identify lateral femoral condyle, tibial plateau, and meniscus (posterior horn)

Probing phase

12. Come back over into medial compartment again following intercondylar notch (supervising consultant to assist with positioning)
13. Insert probe under direct vision
14. Probe articular surface of medial femoral condyle
15. Probe weight-bearing surface of medial tibial plateau
16. Probe and lift medial meniscus (posterior horn)
17. Enter intercondylar notch. Probe anterior cruciate ligament
18. Enter lateral compartment following intercondylar notch (supervising clinician to assist with positioning in figure-of-four)
19. Probe articular surface of lateral femoral condyle
20. Probe weight-bearing surface of lateral tibial plateau
21. Probe and lift lateral meniscus (posterior horn)

Figure 6.1: Diagnostic knee arthroscopy sequence from agreed in-theatre protocol, Appendix E.1, pg. 225

Participant characteristics are presented as frequencies for categorical data, and medians with IQR for continuous data. Non-parametric methods (Mann-Whitney U test) were used to assess differences in distributions for each motion analysis metric between the intervention and control groups at baseline and follow-up on the simulated arthroscopy list and during the in-theatre assessments. A p-value of 0.05 was regarded as statistically significant. Holm-Bonferroni sequential correction was used to generate p'-values for baseline simulator assessment, follow-up simulator assessment, and in-theatre assessments. All statistical analyses were performed in R v3.3.2.²⁸⁴

6.3 Results

Sixty-six surgical trainees were invited to participate. Thirty responded to the invitation and all were eligible. They were randomly assigned to the control or intervention group (15 in each group). The baseline characteristics were balanced between both groups for their age, sex, hand dominance, surgical experience, and simulator exposure (Table 6.1). Of the 30 participants starting the study, 28 completed the study: two participants dropped out of the control group (Figure 2; one left the training programme before the conclusion of the study, and one withdrew because they weren't randomised to their preferred group, the intervention group).

6.3.1 Participant baseline simulated arthroscopy list performance

At baseline, the median number of hand movements to complete the simulated DKA in the intervention group was 308 (IQR 250-426) compared to 311 (IQR 286-424) in the control group. This difference was not statistically significant ($p'=1$; Figure 6.3). There was no statistically significant difference between the two groups in the number of minor movements, smoothness, or time taken to complete the simulated DKA, nor their performance in any metric on any of the SAL tasks ($p'=1$ in all cases; see Appendix E.2).

Table 6.1: Baseline characteristic of participants according to randomisation. Data reported as median (IQR) or n (%)

		Control (n=15)	Intervention (n=15)
Sex	Female	6 (40%)	8 (53%)
	Male	9 (60%)	7 (47%)
Handedness	Left	1 (7%)	0 (0%)
	Right	14 (93%)	15 (100%)
	Ambidextrous	0 (0%)	0 (0%)
Age (years)		26.3 (26.1–27.2)	26.0 (25.8–27.2)
Postgraduate year	FY2	12 (80%)	12 (80%)
	CT1	3 (20%)	3 (20%)
Arthroscopic experience	Observed	1 (0-5)	2 (0-6)
	Assisted/Performed	0 (0-0)	0 (0-0)
Other minimally-invasive surgery experience	Observed	10 (4-15)	10 (4-16)
	Assisted/Performed	1 (0-4)	1 (0-3)
Open surgical experience	Observed	10 (8-25)	20 (9-25)
	Assisted/Performed	3 (0-12)	2 (0-8)
Previous simulator experience	Cadaveric	0 (0%)	1 (7%)
	Animal	0 (0%)	0 (0%)
	Anatomical bench-top	2 (13%)	1 (7%)
	“Box” trainer	8 (53%)	3 (20%)
	Virtual reality	1 (7%)	2 (13%)
Frequency of simulator training	Daily	0 (0%)	0 (0%)
	Weekly	0 (0%)	0 (0%)
	Monthly	0 (0%)	0 (0%)
	‘Termly’	0 (0%)	0 (0%)
	Annually	0 (0%)	1 (7%)
	Less often	3 (20%)	3 (20%)
	Never before	12 (80%)	11 (73%)
Use of simulators	On a training course	7 (47%)	5 (33%)
	As a research participant	1 (7%)	3 (20%)
	Job interview	0 (0%)	1 (7%)
Estimated prior simulator use (mins)		15 (0-60)	20 (0-90)

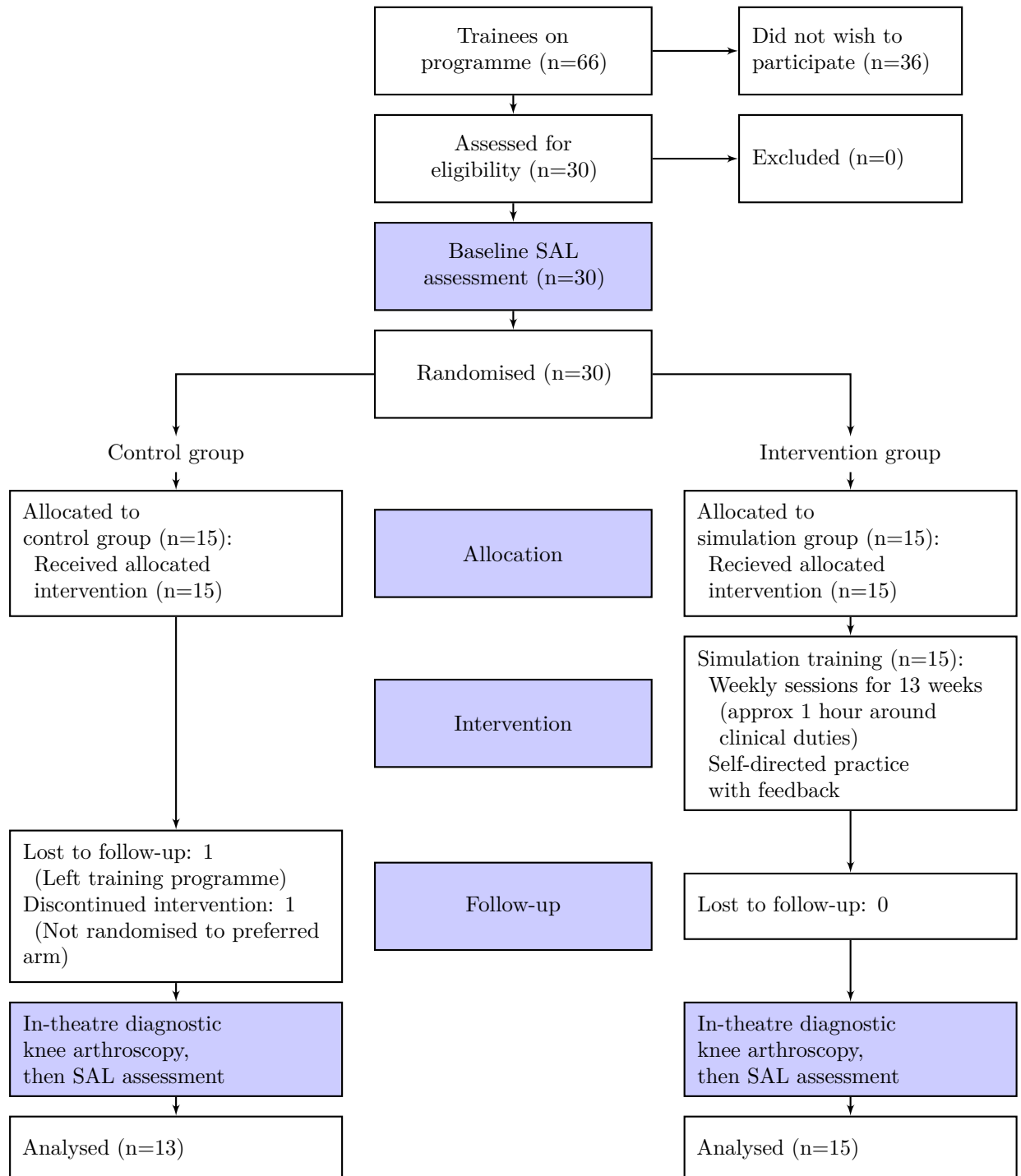


Figure 6.2: CONSORT flow diagram

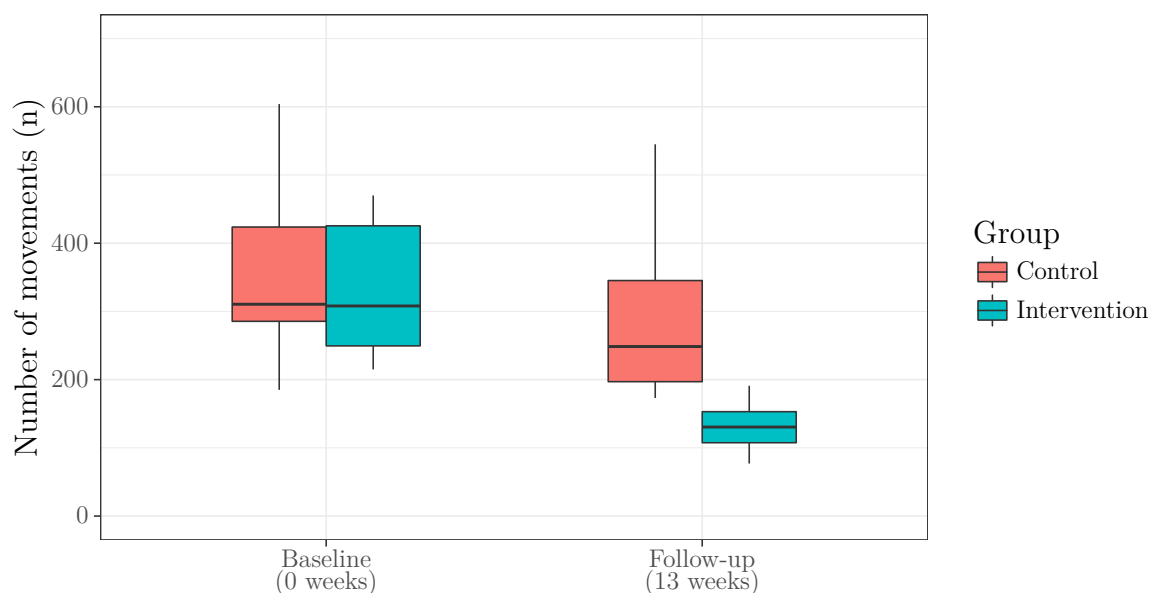


Figure 6.3: Simulated diagnostic knee arthroscopy performance at baseline and follow-up. There is no difference between the groups at baseline, but the intervention group require fewer hand movements than the control group at follow-up.

6.3.2 Participant in-theatre performance

After 13 weeks, during the in-theatre assessment the intervention group consistently outperformed the control group (Figure 6.4). The median number of hand movements to complete a supervised DKA in the intervention was 544 (IQR 465-593) compared to 893 (IQR 747-1242) in the control group. This difference was statistically significant ($p'=0.00048$). The number of minor movements, smoothness, and time taken were also statistically significantly lower in the intervention group ($p'\leq 0.00048$ in all cases; see Appendix E.5).

6.3.3 Supervisor in-theatre performance

All supervising consultants in theatre demonstrated the DKA to the participants on the same patients before their attempt as is usual training practice. There were no statistically significant differences in their performance while they were supervising the control or intervention groups across all the objective outcome measures ($p'=1$ in all cases; Figure 6.5, and Appendix E.5).

6.3.4 Participant:supervisor in-theatre performance ratio

The relative performance of each participant compared to their supervisor on the same case was calculated as a ratio for each metric. This adjustment shows how the participants performed compared to presumed optimal performance in that case. The intervention group required at median 1.9 hand movements (IQR 1.5-2.1) for each supervisor movement compared to 3.3 (IQR 2.2-4.8) in the control group ($p=0.018$). Relative to the supervisor, the intervention group also required fewer minor hand movements, were smoother, and took less time than the control group ($p\leq 0.018$ in all cases; Figure 6.6, and Appendix E.5).

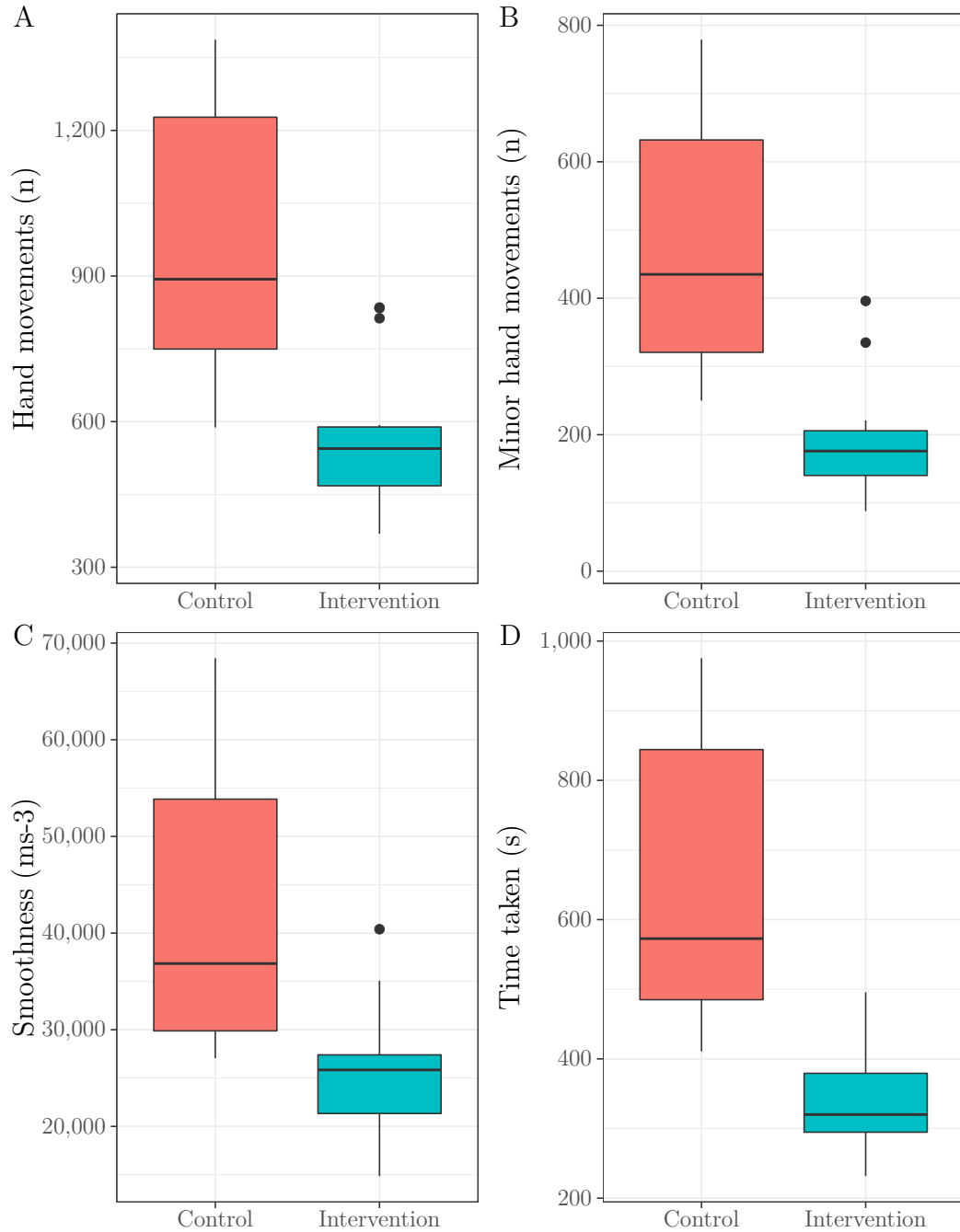


Figure 6.4: Participant performance measured by (A) the number of hand movements, n , (B) the number of minor hand movements, n , (C) the smoothness, ms^{-3} , and (D) time taken, s , to complete a diagnostic knee arthroscopy in the operating theatre after 13 weeks of standard surgical training $\pm$ additional simulation. The intervention group outperform the control group in all objective metrics ($p' < 0.05$ in all cases).

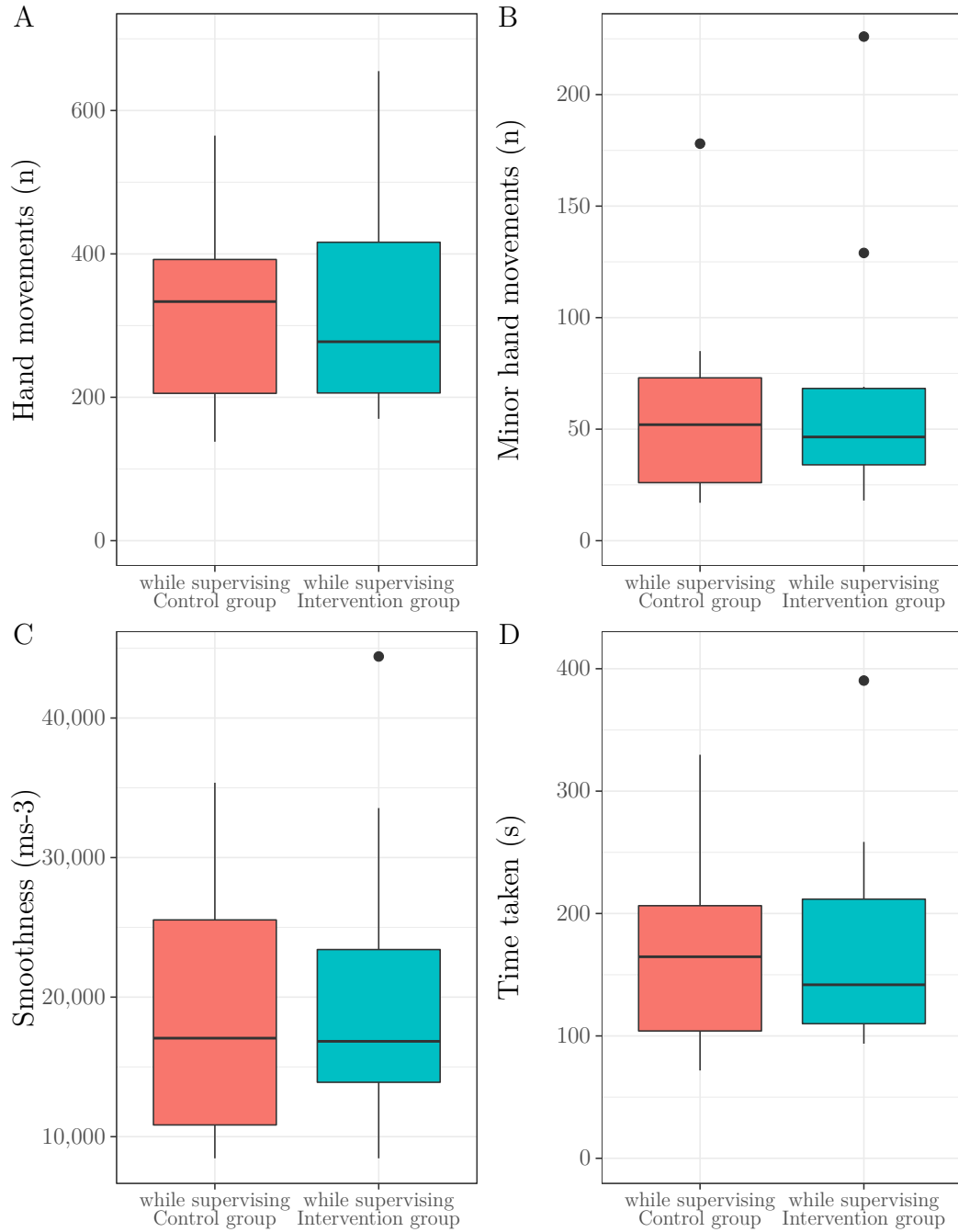


Figure 6.5: Supervisor performance measured by (A) the number of hand movements, n , (B) the number of minor hand movements, n , (C) the smoothness, ms^{-3} , and (D) time taken, s , to complete a diagnostic knee arthroscopy in the operating theatre when supervising the control and intervention groups. There is no statistically significant difference in performance by any of the objective metrics ($p > 0.05$ in all cases).

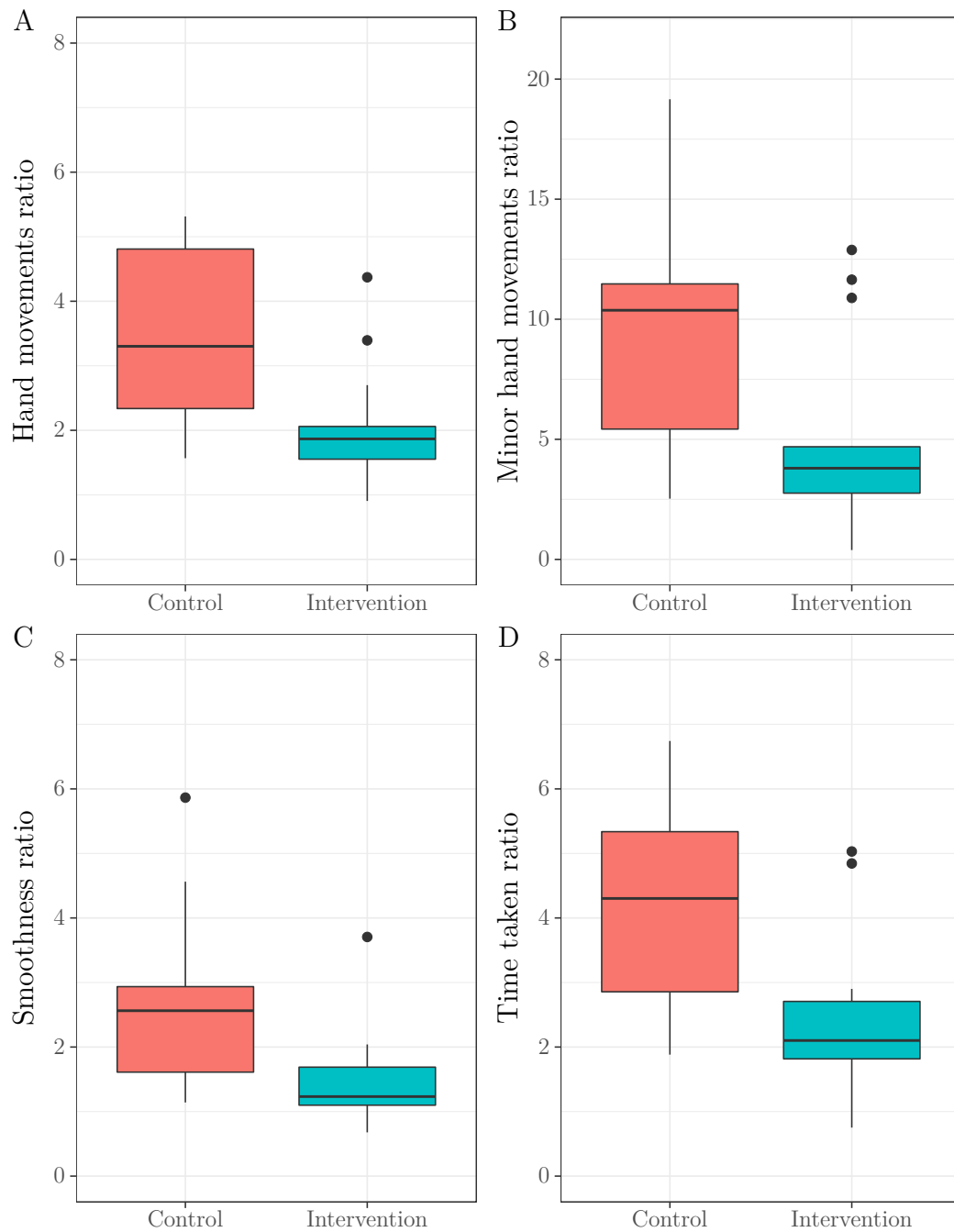


Figure 6.6: Ratio of participant:supervisor performance during the same case as measured by (A) the number of hand movements, n , (B) the number of minor hand movements, n , (C) the smoothness, ms^{-3} , and (D) time taken, s , to complete a diagnostic knee arthroscopy in the operating theatre after 13 weeks of standard surgical training $\pm$ additional simulation. The intervention group perform closer to their supervisors performance on a given case than the control group for all objective metrics ($p' < 0.05$ in all cases).

6.3.5 Participant follow-up simulated arthroscopy list performance

At follow-up, the intervention group generally performed better than the control group at the SAL tasks. The median number of hand movements to complete a simulated DKA in the intervention group was 131 (IQR 108-153) compared to 249 (IQR 197-345) in the control group ($p'=0.028$; Figure 6.3). There were statistically significant differences between the two groups in the number of minor movements, smoothness, and time taken to complete the simulated DKA, and their performance in the SAL tasks ($p'\leq 0.03$ in all cases), except for the number of hand movements during the Maze and Flat ring remove tasks ($p'=0.068$ in both cases; see Appendices E.1-E.4).

6.3.6 Participant training activities during study

Baseline surgical experience is recorded in Table 6.1 indicates similar arthroscopic, minimally-invasive and open surgical experience, and simulator exposure, between the groups at baseline. The participant training activities during the study period is summarised in Table 6.2. The median number of simulation sessions attended by the intervention group was 10. There was no significant difference in the number of arthroscopic, minimally invasive or open surgical cases that the groups attended ($p>0.05$ in all cases). Two participants in each group attended a training course which involved simulation, spending between 30 and 60 minutes in total on the simulator.

6.4 Discussion

6.4.1 Implications of the study

This randomised controlled trial demonstrates the first evidence that performance can be objectively measured in theatre to show practice on simulators during clinical training objectively improves technical skills acquisition and operating performance of trainee surgeons. Participants who received simulation training performed significantly better than those receiving standard training only, based on objectively

Table 6.2: Participant training activity during the study, according to randomisation. Data reported at median (IQR) or n (%)

		Control (n=13)	Intervention (n=15)	
Study simulation sessions attended		N/A	10 (8-12)	
Arthroscopic experience	Observed	0 (0-1)	2 (0-2)	p=0.19
	Assisted/Performed	0 (0-1)	0 (0-1)	p=0.88
Other minimally-invasive surgery experience	Observed	0 (0-0)	0 (0-1)	p=0.053
	Assisted/ Performed	0 (0-0)	0 (0-0)	p=0.13
Open surgical experience	Observed	2 (0-11)	4 (0-16)	p=0.66
	Assisted/Performed	0 (0-5)	0 (0-2)	p=0.13
Use of simulators	On a training course	2 (15%)	2 (13%)	p=1
	As a research participant	0 (0%)	0 (0%)	p=1
	Job interview	0 (0%)	0 (0%)	p=1

recorded surgical hand dexterity measures during real-life knee arthroscopy. Since diagnostic arthroscopy is repeatable, the participants' performance could also be normalised to the expert supervisor's performance under the same conditions to account for patient factors. Normalised to supervisors performance, the intervention group required fewer hand movements, fewer minor movements, less time, and were smoother than the control group, but they did not perform as well as their supervising surgeons. The study's strengths include its use of objective metrics; study design to assess the transfer of training to intra-operative performance; self-directed simulation with knowledge of results; the ability to measure supervisor performance to assess case mix and standardise participant performance; the pragmatic provision of simulation training which could be implemented around normal clinical training; and comparing additional simulation training to a normal training programme rather than the absence of training.¹⁴⁴

Elsewhere in the literature there are implications that simulation training leads to improved intra-operative performance.^{144,145,327} Bench-top and virtual-reality

simulators have used motion analysis^{87,276} to show that simulation training improves simulated performance,^{197,204,234,247,328,329} but until this study, any intra-operative comparisons have relied on subjective scales.^{144,145,185,186} Global rating scales and related tools were previously the only feasible options to be used for assessment in theatre, but they can be time-consuming,^{198,330} have ceiling effects,^{250,262,280} and are fundamentally subjective.^{11,86,87} Global rating scales do have an advantage over motion analysis in that they allow a subjective assessment of the end product of surgery, but because these scales summarise a subjective assessment of performance to an integer they imply that they are objective³³¹ which leads to controversial statistical analysis.¹³⁰ Intuitively, metrics such as patient outcomes (survival, complications, PROMs) and surgical process outcomes (blood loss, operative time) seem useful for assessment, but these are surrogates for intra-operative performance and can be affected by factors beyond the operating surgeon.²³ This is the first study to objectively show the transfer of surgical simulation training to intra-operative performance using objective performance metrics relevant to intra-operative activity.

Simulation training has been recognised by Nousiainen et al¹⁵³ to increase costs per trainee by 15 times, and Acton et al¹⁵⁴ report simulation training has required a 350% increase in teaching faculty but despite extra staff 93% of programme directors still report an increase in faculty workload. Competency-based simulation training can be both resource and faculty intensive. Though a formal cost analysis has not been calculated in this study, demonstrating that simulation training leads to intra-operative behaviour change is an important step in justifying those investments. Though further work is needed to understand how cost-effective investment in simulation is, it is anticipated that the self-directed deliberate practice with proximate feedback described in this study would be less burdensome on faculty. Simulation resources could be used more intensively if staff didn't need to be present during all training to provide feedback, and trainees would not need to attend simulation in typical working hours which detracts from service commitments.

Transfer of training is often misinterpreted in the literature. Practice in one simulation resulting in improved performance in another is sometimes reported as

‘transfer of training’.^{106,113,187,189,283,327,332} This confusion in terminology may be a genuine or disingenuous semantic error and has been highlighted in the by others in the field.³³³ Such studies indicate generalizable elements and the concurrent validity of the simulations, but they do not show ‘transfer validity’ i.e. a transfer of simulated skills to intra-operative performance, or Kirkpatrick Level 3. Some of these studies use live or cadaveric animals^{286,327} for ‘in-theatre’ assessment: although arguably more realistic than bench-top or virtual-reality simulators and preferred by trainees,^{136,137} these are still surgical simulations and cannot fully represent surgery on live patients.

How does simulation work? One assumption is that it permits practice of generalizable psycho-motor skills, but additional simulation training may not directly affect technical skills. Surgical technical skills have traditionally been learned in theatre. Gallagher notes the importance of cognitive load during surgery having sufficient capacity to acquire new skills,^{50,105} while Hill highlights the ‘hidden curriculum’ and the role of building a network and “gaining participation in the surgical world”.⁴⁸ If simulation increases a surgeon’s confidence to help navigate the hidden curriculum of surgery or reduces cognitive loads, this could result in privileged access to teaching in theatre or could give trainees the capacity to learn more effectively in theatre. The cause of the observed improvement in intra-operative performance at follow-up could be that simulation facilitates an increased rate of skills acquisition by boosting theatre attendance of “unskilled and unaware” trainees,¹⁷⁶ rather than shifting the initial, more dangerous, stages of training out of the operating theatre into the skills lab. Any confidence boosting effect of simulation was not directly assessed in this study, but no difference was noted between the groups in the rates of theatre attendance during the study. This suggests the impact of simulation training was the main factor driving technical skills differences in theatre, and supports the evidence from the skills lab where practice improves simulated performance.

Deliberate practice and knowledge of results are influential factors in the efficient acquisition of complex motor skills,^{100,107} but deliberate practice has had relatively

little application in simulation.³³⁴ Simulation training studies commonly rely on simple repetition or ‘training to proficiency’ on the same model.^{129,144,145} Because proximate objective motion feedback was available, this study was designed so that simulation was led by the participants who chose which tasks to practice based on their feedback from the motion analysis system. In practice sessions, the intervention group participants could attempt the SAL tasks and compare their performance to novices, trainees and consultants using data from the construct validation study in Chapter 4. As each task is designed to test different arthroscopic sub-skills, participants could see where they were weaker or stronger, and use the feedback to focus their training efforts. The motion analysis system allowed them to view their own progress over time to see which practice strategies were effective at improving their performance.

6.4.2 Limitations

It would have been preferable to record the baseline intra-operative performance for both groups to ensure they were matched at the outset and to compare longitudinal performance changes, but this was logistically impractical given the number of participants and the available operating lists. However, the SAL has been validated to distinguish participants based on surgical experience (Chapter 4) so the baseline simulator performance indicates the groups were well-matched initially as does their baseline demographic information (Table 6.1). After 13 weeks, the intra-operative differences are mirrored by the differences in follow-up simulator performance (Appendix E.2). This is a pragmatic alternative to intra-operative assessment at baseline.

This study focused on improving the arthroscopic skills of relatively junior trainees, who some feel should not yet be learning such advanced techniques³³⁵ or that the findings will not translate to high-level trainees. This group were chosen as they potentially have the most to learn. There is evidence that the early stages of learning new procedures are associated with most risk of harm to the patients^{25,112} so they benefit from shifting the learning curve out of the operating room and

into the skills lab.¹¹⁶ Although we show intra-operative motion analysis can assess changes in technical skills between the groups in this study, other performance metrics such as global rating scales and task-specific checklists have shown ceiling effects^{250,280} and further work is needed to validate wireless motion analysis to assess performance variations in more experienced trainees which may be more subtle.^{91,103}

The diagnostic knee arthroscopy procedure performed may be seen as a simple operation, which is no longer indicated in and of itself.¹⁴⁸ Although it is unlikely to directly affect patient outcome when compared with some therapeutic interventions,³³⁶ it is the most suitable task for assessing early trainees from a patient safety perspective. DKA is common to all arthroscopic procedures at the outset to confirm examination and radiological findings, and at conclusion to confirm the adequacy of the procedure performed. It utilises generalisable arthroscopic skills such as navigation and triangulation, and the sequence is standardisable sequence which is important for assessment. Meniscectomy, a plausible alternative, would need to account for the variety of tear morphologies and the procedure would not be repeatable.³³⁷ Staple arthroscopic training procedures such as meniscectomy and sub-acromial decompression are currently the subject of some controversy.^{150,336} If these procedures become less common, trainees will have fewer opportunities to learn in theatre, so simulation of 'simple' procedures may be required as a bridge to more complex arthroscopic surgery. Additionally, the diagnostic knee arthroscopy is repeatable by trainer and trainee allowing a 'performance ratio' to be calculated, which is a key to advantage in this study.

The calculated performance ratios have not previously been validated. These are new metrics that rely on the assumption that a supervisor's performance is 'optimal' for a given patient. It is not unreasonable to assume that a consultant supervisor will perform optimally during this most fundamental arthroscopic procedure. All supervisors understood their role was to provide an idealised example to reference the participant performance against, and none requested a repeated attempt to improve their performance. Using a simple ratio implies that the relationships between performance, patient factors and experience are linear, although this

is unlikely. There is evidence that the relationship between performance and experience is not linear as discussed in Chapter 5, but the role of patient factors is not well understood. As the field develops, it might become apparent that another mathematical function to reference performance is more appropriate, but for this study it is logical to attempt to adjust for patient factors and a ratio is simple enough as not to obfuscate. Notably, performance measured using the unadjusted metrics also showed significant differences between the groups.

This study was not able to identify optimal training intervals or total simulation time, or define a competence level to achieve in the skills laboratory before operating in theatre. These are important issues but this study was set up to allow pragmatic simulation training around day-to-day clinical duties. As the training was self-directed and trainee-led, training intervals and total simulation required were not variables that could be controlled. Even with this flexibility, fewer than 80% (10/13) of planned sessions were attended (Table 6.2). The optimisation of simulation in terms of the logistics of simulation programmes, the surgical tasks, the stage of trainees, and levels of pre-patient competence can now be more effectively defined with the benefit of validated objective performance metrics which can be used in and out of theatre.

6.4.3 Summary

This study provides evidence for the first use of wirelessly recorded motion metrics in real surgery to objectively demonstrate the transfer of simulation to intra-operative arthroscopy performance. The metrics described are an objective, quantitative and proximate performance measure that can be used as an adjunct to traditional methods in surgical training, and allow direct comparison of simulator and in-theatre performance. Simulation training can be effectively provided around clinical duties and as a self-directed intervention reducing the requirements on training faculty. Appropriately validated, these metrics could translate to other surgical specialties. This novel objective assessment method offers opportunities to link simulation training and intra-operative performance to patient outcomes after real

6. Randomised controlled trial to objectively assess the transfer of surgical simulation training to intra-operative performance 170

surgery; to define competence levels for pre-patient training, career progression and revalidation; and potentially for talent identification and training selection.

A surgeon should limit his personal service strictly to those fields in which he is a master.

— Dudley Allen

7

Discussion

Contents

7.1	Summary of main findings and conclusions	171
7.1.1	Contributions to the field of study	174
7.2	Clinical relevance	175
7.3	Limitations	176
7.4	Recommendations for future work	180

7.1 Summary of main findings and conclusions

The review of the literature in Chapter 1 showed that surgery has evolved dramatically in the last 200 years. The roots of the manual discipline of surgery are preserved in the word itself, but as the technical scope of surgery increased the primacy of technical skill has been ceded to other skill domains. At the same time, care has become more patient-centred and evidence-based, previous training practices and working patterns are now unacceptable, and surgeons are under greater scrutiny. Advancing arthroscopic techniques present new technical skill challenges, but current trainees have not grown with the discipline as it develops like their predecessors, and unlike open surgery there are no day-to-day analogues where surrogate skills are acquired. Simulation training has been proposed as the

pedagogical, ethical, and practical solution to changes to surgical technical skills training. However, simulation is costly and time-consuming, and the evidence of its impact on intra-operative was subjective and inferred from other fields. Therefore, this thesis proposed to answer the question:

Can technical skills practice through arthroscopic simulation lead to objective improvements in intra-operative arthroscopic surgery performance?

The systematic review in Chapter 2 identified how quantitative metrics for technical skills assessments had been applied to arthroscopy (Objective 1) and found 92 arthroscopic studies using quantitative methods to assess technical skills performance. These were grouped into 10 classes and the implications and evidence for each metric class were discussed. Five of the metrics identified had been applied to the operating theatre in 20 studies and none linked performance to patient outcomes. There was only limited evidence from four studies that simulator training led to a change in intra-operative behaviour using three of these quantitative metrics (Global rating scales, Task-specific checklists, and Single-item scales). Though quantitative, all require subjective interpretation by an observer and were not suitable for independent simulator practice. The systematic review showed motion analysis, which had for technical reasons struggled to be implemented in the operating theatre, was a validated method of recording objective quantitative metrics in the skills lab and identified wireless elbow-worn inertial motion sensors as a candidate to provide a suitable objective quantitative metrics of skills assessment for use in a simulated and surgical setting (Objective 2).

The studies in Chapter 3 validated a wireless elbow-worn inertial motion analysis system against a wired wrist-worn electro-magnetic motion analysis system which was commonly used in skills labs but not suitable for the operating theatre. The first study showed that when collecting both the accelerometer and gyroscopic data the counts of hand movements had a strong correlation to the wired device ($r_s=0.91$), although it was more sensitive with a consistent bias recording a mean 28.3 hand movements more than the wired system. The second study showed that the wireless

elbow-worn motion analysis system was accurate and reliable, recording a median of 21 hand movements (IQR 21-22) after 20 stereotyped arthroscopic movements were made along a three-axis rig to control distance travelled and extraneous motion. As such, this wireless system was considered a suitable objective, quantitative method of skills assessment to validate further for a simulated and surgical setting (Objective 2).

The paired studies in Chapter 4 developed and validated a simulated arthroscopy list to allow independent, deliberate practice of the fundamental technical skills associated with arthroscopic surgery (Objective 3). A group of orthopaedic fellows, who had recently completed their training and could consider the T&O curriculum with insight of its final requirements in the context of a modern training environment, selected four generic skills modules for the FAST workstation which they considered most appropriate for the aims of the curriculum to train the skills required for novice trainees. These four modules (Probe, Maze, Flat ring transfer*, and Arc cutting) were combined with an established diagnostic knee arthroscopy exercise and a novel meniscectomy exercise, making the ‘simulated arthroscopy list’. These tasks were assessed with wireless motion analysis metrics for their ability to distinguish between novice (medical student), intermediate (orthopaedic registrar) and expert (post-CCT fellow/consultant) levels of performance in a construct validity study. Four of the eight motion analysis metrics (number of hand movements, number of minor hand movements, smoothness, and time taken) were consistently able to distinguish between experience levels on all tasks except the Maze. These four metrics were felt by participants to be more useful (67-80% vs 53-63% of respondents), represent performance more clearly (83-100% vs 57-70%), and less likely to be uninterpretable as a guide to improve performance (0% vs 7-33%) than the metrics which did not show construct validity. The SAL was felt to improve general and specific arthroscopic skills, to be sufficiently realistic and suitable for novice trainees (97% of respondents).

An important assertion of simulation training is that it is more efficient and can accelerate the safe acquisition of technical skills. In Chapter 5, a randomised

*Run as two separate tasks, first removing, then returning, the rings into position

study was conducted which investigated whether deliberate practice with objective performance feedback would optimise skills acquisition (Objective 4). Practising on the Flat ring transfer module, candidates who practiced with and without feedback both improved their performance significantly according to the number of hand movements, number of minor hand movements, smoothness, and time taken metrics. This study showed a marked impact of deliberate practice on performance, but there was no consistent difference in level or rate of performance between the groups receiving feedback or without feedback. Notably, when performance feedback was removed from the intervention group in the final practice session, there was no deterioration in performance to suggest they had become ‘dependent’ on feedback.

Having shown that the evidence for the transfer of simulated skills to improved operating room is based on subjective assessment, in Chapter 6 a randomised controlled trial was conducted to objectively assess the transfer of training follow practice on the SAL (Objective 5). After 13 weeks, it showed that participants who received self-directed simulation training performed a diagnostic knee arthroscopy in theatre with fewer movements, more smoothly, and in less time than those who had standard surgical training alone ($p' \leq 0.018$ in all cases). Their supervisors also wore wireless motion sensors, and there was no difference in their performance while they were demonstrating the procedure to the intervention or control group, implying the performance differences between the control and intervention groups was not due to patient differences. By assuming that supervisors would be able to perform this basic procedure as well as could be done for a given patient, participant performance on the same patient was presented as a ratio of the ‘optimal’ supervisor performance to account for patient differences. Again, this showed that the intervention group were quicker, smoother and required fewer movements than the control group.

7.1.1 Contributions to the field of study

This thesis contributes to the field of study in several ways:

- The systematic review identified a range of quantitative metrics in use to assess arthroscopic performance, but that the evidence for simulation improving

intra-operative performance is based on subjective assessments and there is no evidence that it improves patient outcomes.

- Wireless elbow-worn inertial motion sensor analysis produces motion metrics which show construct validity, can be used independently to guide training, and is the first system that records a surgeon's motion to be used in the operating theatre.
- The randomised controlled trial in Chapter 6 is the first to use objective, quantitative and proximate metrics to report the transfer of simulated arthroscopy training to improved intra-operative performance.
- These studies assessed the impact of self-directed deliberate practice on intra-operative performance, to understand how independent simulation can be effective to reduce the burden on training faculty and the associated costs.

7.2 Clinical relevance

The work presented in this thesis has important clinical implications for the use and assessment of simulation in orthopaedic training:

- Technical performance can be measured in an objective, quantitative manner, proximate to the event in both the skills lab and operating theatre by using the wireless motion analysis metrics. These can provide useful feedback for deliberate practice which may, in the right setting, be beneficial to the acquisition of expert performance.
- These objective, quantitative metrics will provide an alternative perspective on technical performance to complement, not replace, the current subjective, quantitative (e.g. GRS, TSCL) and subjective, qualitative (e.g. direct supervisor feedback) assessments. As they can be used independently, this will alleviate the time and cost burden on training faculty, to make simulation training more readily available.

- Simulation training has been shown in this thesis to objectively improve intra-operative performance. Much research effort is currently expended on validation studies, and this should lead to a change in research strategy in the simulation community to focus on addressing how to optimise the impact of simulation on intra-operative performance, and whether this results in improved patient outcomes.
- The rapid performance improvements in the learning curve study (Chapter 5), beyond the ‘consultant level’ performance derived from the construct validation study (Chapter 4), highlight that performance does not equal learning. This has important implications for how simulation is applied in pre-patient training. Inferences from simple simulator tasks about preparedness for intra-operative procedures and decisions to transfer surgical responsibility to trainees should be supported by other evidence.

7.3 Limitations

One of the necessary elements of this thesis is the ability to objectively, quantitatively measure intra-operative activity and relate these measurements to surgical ability. Though the studies in this thesis identified the knowledge and technology gap in assessing the transfer of simulated skills to the operating theatre (Chapter 2), validated of the motion analysis system against common alternative system and when in use by practitioners of differing level of ability (Chapters 3 & 4), there are limitations in this approach. The motion analysis system may be seen as a ‘black box’ with concerns about the nature of the metrics being used to draw conclusions about performance. Performance was also framed in the context of skill stratification in the construct validation studies in Chapter 4 which relies on assumptions about the groups defined (in this case, Expert, Intermediate and Novice). In addition to caveats in interpreting performance based on the thresholds used to define the experience of the trainees, the tasks chosen are relevant, and in particular the maze exercise which was included despite its poor quantitative data from Chapter 4.

The motion sensor hardware used is commercially available, but the algorithms used to convert the raw accelerometer and gyroscope signal into the motion analysis metrics is intellectual property of McLaren Applied Technologies. It is important to recognize the ‘black box’ nature of the way the results are generated when we cannot review exactly how the outputs relate to the inputs. Such black boxes and opaque measures of reporting outcomes hinder transparency, replication of results, and the broader use and acceptance of the method by others in the field, though even transparently published outcome measures may be financially restricted by licensing and so be used less widely or replicated less often than would otherwise be the case. Although we do not know the exact values for the signal filters, transformations, and thresholds used when processing the motion data into the metrics generated, these were based on published signal processing algorithms.^{220,223,276} These related published algorithms have shown construct validity and concurrent validity with other methods of assessing technical skills.^{231,338} When recording movements, physiological tremor and experimental error are not intended movements by the surgeons but are recorded and add ‘noise’ to the signals to be processed. Changing the spatial and temporal resolution of the motion analysis devices will affect the amount of noise detected, and this can be removed or smoothed with signal processing techniques.²⁷⁶ Different degrees of smoothing and other signal processing methods, as well as the raw signal itself, will result in different values for motion metrics from different algorithms though the principles behind them remain similar. The metrics generated may not be directly interchangeable between systems, but it may not be critical that the numbers generated represent the ‘true number’ of hand movements, for example, if the values reported are reliable, accurate, and of an appropriate order to be interpreted by the user. Proprietary algorithms are valuable to the companies that create them and are especially common in VR simulation systems. One should be circumspect regarding the exact values for motion metrics produced and how interchangeable these are between simulation systems or environments, but their use can still be appropriate if they are shown to have concurrent validity with existing metrics and display construct validity.

Stratifying practitioners with different experience levels is a useful way to assess a simulation against a standard: although you do not know the exact performance differences, the direction of correlation is anticipated based on the differences in time in surgical practice. The studies in this thesis used three groups: Novice (little to no surgical experience) and Expert (high surgical experience) can be well-defined and made quite uniform, but the Intermediate group is heterogeneous. It covers a period of significant learning and range of ability. It represents a transitional period and does not constitute a single experience level in the same way that medical students are total novices or established consultants are capable of independent practice. The three groups are defined by the experience thresholds between them: these three groups are common in simulation literature and pragmatic for recruitment,^{127,194,223,231,255} as are studies with the even less granular two groups (Surgeons/Non-surgeons),^{91,202} while more granular divisions are infrequent in the literature.^{84,205} The thresholds between the three groups used in these studies were chosen because they represent the levels identified in the Trauma and Orthopaedic curriculum,⁸⁹ though this document was designed as a guide for progress through training and not a method to rank trainees. Price et al have used more granular descriptions of skills, breaking the Intermediate group further into three sub-groups (junior trainees, registrars, and fellows) or a continuous scale of experience by using the number of arthroscopies performed⁸⁴ and have also shown continuous improvement between each experience level. However, this requires the introduction of new thresholds for experience which would be arbitrary and not representative of the thresholds or requirements of the T&O curriculum. Additionally, the studies performed would have required more participants for increased granularity of experience levels which is impractical (our deanery is small and has only 2-3 trainees at each year from ST3 to ST8, so recruiting sufficient numbers would require access to other deaneries or studies lasting a number of years). The Intermediate group is not a single stage but a transition between the Novice and Expert states, and where possible recruiting more participants would be preferable to increase granularity. Where this is not possible, the mean represents a point between Novice and Expert

performance but cannot be used to judge those in the transition directly and using it as a reference for comparisons between participants in that Intermediate group transition phase will be of limited value.

The Maze task was included despite the quantitative data generated in Chapter 4 showing the task did not distinguish participants based on experience. This is partly because the simulated arthroscopy list being developed was attempting to serve two purposes: to be both an environment to train in, and an environment to assess performance. The construct validity study showed that the Maze task was unable to distinguish between participants of different skill levels which hindered its purpose to assess performance, but it was still included in the final simulated arthroscopy list. This is because the Maze task still contributed to the other purpose: as an environment to train on, it was felt to be one of the top 3 most useful training tasks for novice trainees by all participants in the task selection study and was the only task specifically designed to address the “tracking a moving target” arthroscopic skill identified by FAST curriculum.³¹⁵ Though including the Maze task identified a tension between the role of the SAL in practicing skills, where it provides a subjectively useful practice tool and uniquely targets an otherwise absent skill domain, but not for assessment, where the task could not distinguish participants based on experience, the participants in the intervention group in the transfer validity study (Chapter 6) who practised regularly on the Maze task showed an improved performance in the Maze task compared to their own baseline performance and that of the control group (see Appendix E.2). This implies that the Maze task may be useful as an assessment measure in some circumstances. It may not be sensitive to detect performance differences at a single instance but these reveal themselves with practice and thus the Maze task may be suited to longitudinal assessments. More studies are required to understand the learning curve dynamics of all of the tasks in the SAL, as highlighted by the findings of Chapter 5.

7.4 Recommendations for future work

This thesis examined the behavioural impact of simulation training on intra-operative activity. Questions remain unanswered about surgical simulation, and new questions stem from the findings of this thesis.

The intuitive link between technical skills performance and patient outcomes remains difficult to demonstrate. Partly this has been due to objective performance metrics and case mix factors, which this thesis has considered, but surgery is still a ‘team sport’³³⁹ with many parties contributing to patient outcomes; outcomes which often take longer than a surgeons training to become apparent. To investigate how patient outcomes are affected by changes in performance due to simulation training, a randomised controlled trial would need to identify an procedure where the ‘critical surgical element’ had been confirmed,³⁴⁰ which is performed in sufficient numbers by trainees, with an appropriate simulator available to train on. The choice of outcome measure should be considered carefully: this thesis has stressed the ultimate importance of patient outcomes, and other surgical studies of technical skills have attempted to link these to PROMS, but these are not the only outcomes in a healthcare system. Complications are associated with harms and costs, but are rare and need to be analysed appropriately.¹²¹ Other surgical parameters which have been linked to PROMs or treatment failure such as implant position³⁴¹ could be used as surrogates. Even non-clinical factors, such as the cost and time that trainees add to surgery are important when evaluating simulation training.^{28,29} It is important to recognise that the link between technical skill and patient outcomes may be weak, difficult to assess, or mitigated by other team members, and that indirect improvements to the system as a whole are valid outcomes of interest. Understanding how simulation training affects patient outcomes or surgical processes compared to traditional training alone would allow programmes to make informed, evidence-based decisions about the cost-effectiveness of simulation. As a consequence of the work in this thesis, a study to investigate the impact of simulation training on patient outcomes has been proposed and funded.

The learning curve study in Chapter 5 identified the difficulties in the process of learning curve assessment in surgical training, and highlights the critical difference between performance and learning. Understanding the ‘dynamics’ of the learning curve in surgery will become increasingly important if simulation training becomes more widespread. Work will need to be done to understand how we can assess that a trainee is safe to transition from the pre-patient to patient environment¹¹⁶ if simulated performance improves faster than intra-operative performance. The study in Chapter 6 showed that self-directed simulation training does objectively improve intra-operative performance, but makes no assessment of how optimally to deliver that training. Thirteen weeks (the length training in the transfer of training study) would be too long for many junior doctor training rotations, yet the balance of evidence still suggests distributed practice is more effective than ‘boot camps’ which are expensive, take trainees away from service provision, and have poorer retention.³⁴² Understanding and optimising the learning curve will contribute to making simulation more cost- and time-effective, and can also be applied to intra-operative training and competency-based progression.

Motion analysis metrics were identified in Chapter 2, validated in Chapters 3 and 4, and demonstrated their effectiveness intra-operatively in Chapter 6. Part of their advantage is that they are proximate and interpretable by the trainee, permitting independent deliberate practice. In a research setting to understand simulation training they have been useful, but further research will be needed to understand what their impact will be if more widely available, or applied to other settings such as high-stakes assessment. Outside the research setting for simulation training, if motion metrics are used for decision-making about the transition from pre-patient to patient facing training they risk losing their value as a metric[†], as was seen with ‘unbundling’ of procedures when surgical logbook data was used for trainee assessment and progression.^{122,123} People will respond to the incentives provided, but not always in the way expected. In high-stakes assessment such as selection to surgical training or professional accreditation, the pressures and rewards to ‘game

[†]Goodharts’ Law: “When a measure becomes a target, it ceases to be a good measure”³⁴³

the system' are greater. Chapter 5 showed that performance on a task can improve rapidly with practice, and a fair admission process to surgical training should not be determined by a task that candidates could pay for privileged access to e.g. on courses, or by direct purchase of simulators. Considering simulation assessment for accreditation, pilot studies show sensitivity and specificity to detect qualified surgeons is too low to be used independently.^{25,91,259} A significant body of work is necessary to understand how objective, quantitative data can be used accurately and fairly to support decision-making with subjective and qualitative data. Figure 7.1 provides an everyday example of how various types of information are used to make decisions, and Figure 7.2 shows that we need to identify appropriate metrics for a holistic approach to the decisions we make about surgical training.

Arthroscopic surgery was chosen as a test case in this thesis due to the specific psycho-motor challenges, and the lack of day-to-day analogues to practice on. But in open orthopaedic surgery, the nature and range of movements is different: for example, repeated pronation/supination from a screwdriver, or sudden acceleration/deceleration of a mallet. This may invalidate some motion analysis features, and others may become more useful for assessment. Beyond orthopaedics, arthroscopy shares principles with other forms of minimal-access surgery, and laparoscopy, cystoscopy, and endoscopy are also suitable environments to investigate how simulation can be used to assist training and affect intra-operative performance and patient outcomes. Investigating objective assessment of simulation training in other forms of open and minimal-access surgery is needed for further validation and to justify broader investment in simulation training which would be unsustainable if benefits were limited to one technique in a single surgical subspecialty.

Finally, this thesis has focused on the behavioural aspects of acquiring technical skills which was the first step in the knowledge gap regarding surgical simulation training. Behavioural change is expected to result in clinical change in the form of improved patient outcomes, with further work to this end proposed above. However, behavioural changes are also associated with structural and functional changes in the brain. Learning complex visuo-spatial and psycho-motor skills have been

observed on functional magnetic resonance imaging (fMRI) in diverse populations such as pilots,³⁴⁴ taxi drivers,³⁴⁵ musicians³⁴⁶ and dancers.³⁴⁷ Though their actual tasks are different, these groups are similar to surgeons because they represent the results of expert performance achieved through deliberate practice “vastly superior to that of the general population”.⁹³ In surgery, experience-related differences in neuroanatomical correlates of technical skill have been recorded during simulated surgical knot-tying using near-infrared spectroscopy³⁴⁸ and fMRI,³⁴⁹ and on fMRI using surrogate spatial tasks,³⁵⁰ but there are no longitudinal training studies.³⁵¹ However, linking behavioural to neuroanatomical features in task-based fMRI during more complex simulations than simple knot-tying is very challenging (though an MR-compatible laparoscopic simulator does exist³⁵²), and without the objective quantitative performance metrics identified in this thesis, surgical skills fMRI studies have relied on simulated performance or prior experience for behavioural correlation.

Structural and functional brain changes are an objective, quantitative metric associated with technical performance, and the associated brain changes have been proposed to assess learning and skills acquisition “beyond the attainment of technical proficiency benchmarks”³⁵¹ which was identified in the learning curve study in Chapter 5. Functional MRI features of technical skill acquisition may help objective assessment and assessment of true learning for selection processes. Based on the work presented in this thesis, fMRI analysis is underway to identify for the first time the longitudinal changes associated with novices acquiring surgical technical skills using correlation to objective, quantitative intra-operative behavioural data.

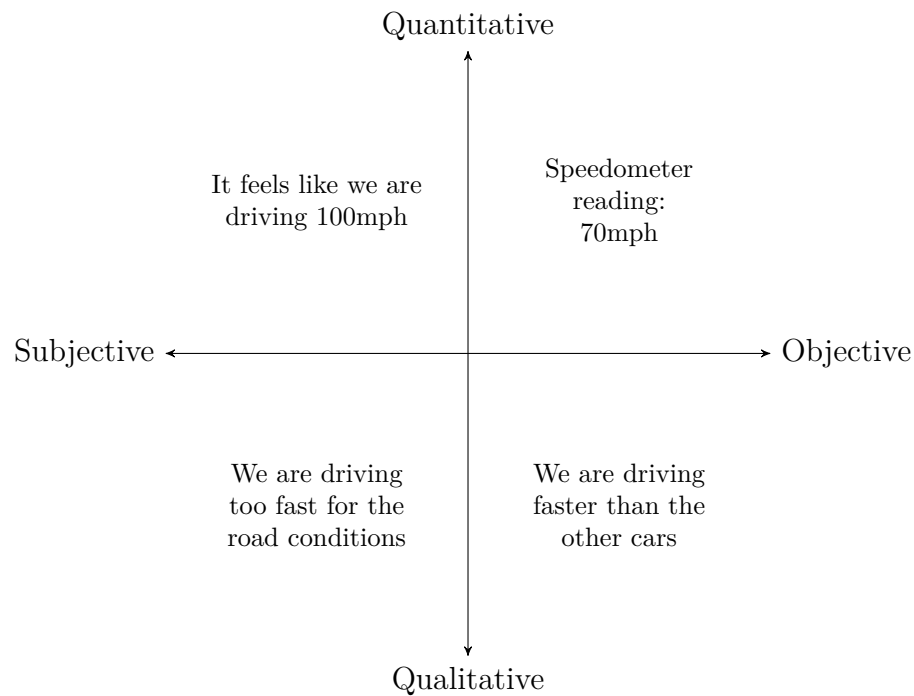


Figure 7.1: Subjective, Objective, Quantitative, Qualitative (SOQQ) plot: Example of how decision-making is informed by a complete understanding of the subjective, objective, quantitative, and qualitative information available for an everyday skill

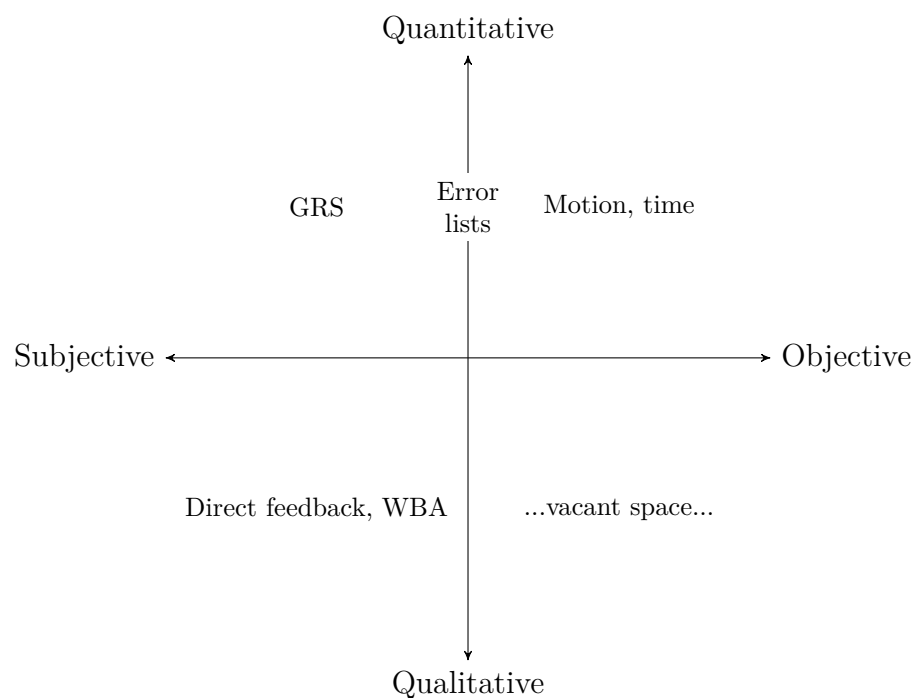


Figure 7.2: Subjective, Objective, Quantitative, Qualitative (SOQQ) plot: Surgical technical skills information available for decision-making about performance during training, assessment, and beyond

Appendices

I am a brain, Watson. The rest of me is a mere appendix.

— Sherlock Holmes in *The Adventure of Mazarin Stone*



General appendices to methods for simulation studies

Contents

A.1	Arrangement of the skills laboratory	186
A.2	Equipment and instruments	187
A.3	Ethical approval	188
A.4	WASP motion analysis system	188

The studies in Chapters 3-6 of this thesis were conducted in the Oxford Orthopaedic Simulation and Education Centre skills laboratory, part of the Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, a department of the University of Oxford. The following appendices are generally applicable to studies conducted in the skill laboratory.

A.1 Arrangement of the skills laboratory

For the studies in Chapters 3-6, simulations were conducted in the skill laboratory in a standardised fashion. The simulation environment was arranged as shown in Figure A.1.

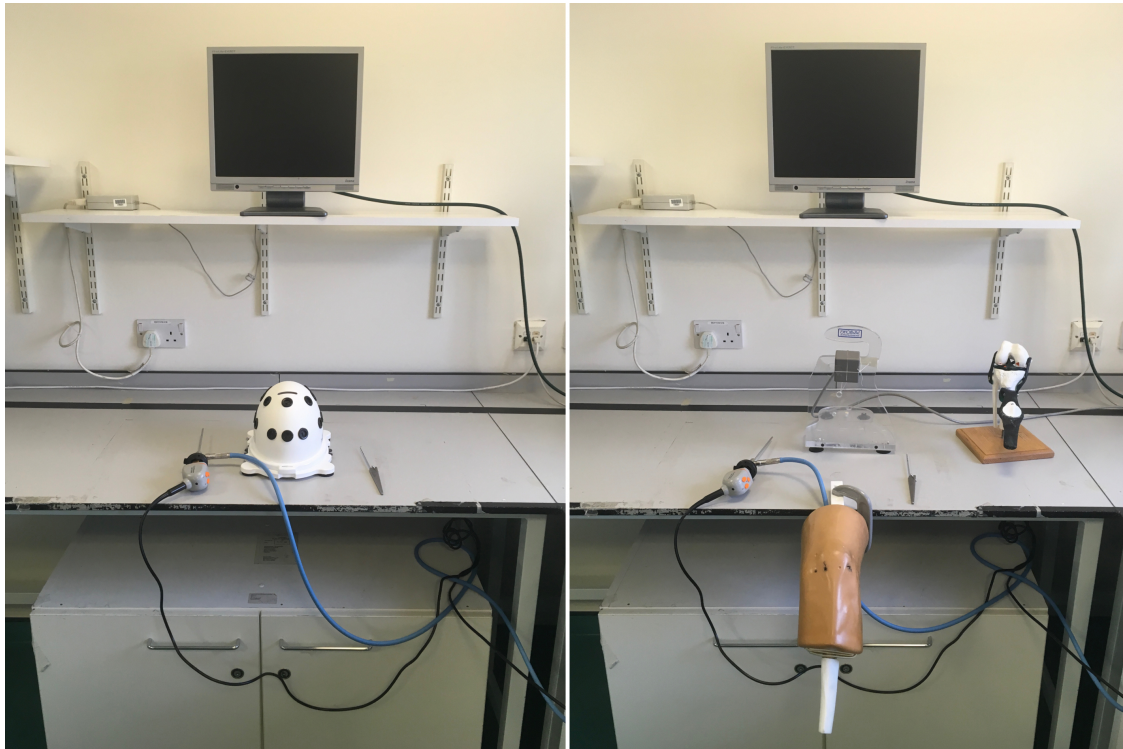


Figure A.1: Skills lab set-up for generic (left) and anatomic (right) bench-top simulation. The Iiyama 15-inch TFT monitor is raised on a shelf at approximately eye-level directly behind the simulator. The simulator is either placed on the table (left, FAST workstation) or attached to the table edge (right, anatomic knee simulator). Arthroscopic equipment to be used in the task is placed on the table before the exercise begins. On the right image, note the position of the grey electromagnetic source behind the simulator (required when using the Patriot wired, wrist-worn motion sensors) and the skeletal knee model (to demonstrate the location of the points to probe during the diagnostic knee arthroscopy exercise).

A.2 Equipment and instruments

The equipment used in the skills laboratory is clinical arthroscopic surgical equipment either decommissioned from the Nuffield Orthopaedic Centre, or received as an educational grant from Smith and Nephew (London, UK). The 30° arthroscope, cannula and 460H camera (Smith and Nephew, UK; Figure A.2a) were connected to the 300XL fibre-optic light source and 360 3-CCD camera image management system (Smith and Nephew, UK; Figure A.2b). Images were displayed on a non-clinical TFT monitor (Iiyama, Tokyo, Japan) as shown in Figure A.1.

The instruments used during the simulations were genuine arthroscopic instruments received as an educational grant from Smith and Nephew (Figure A.3).

Table A.1: Motion analysis metrics available from WASP client in these studies, compared to those reported in Kirby et al.

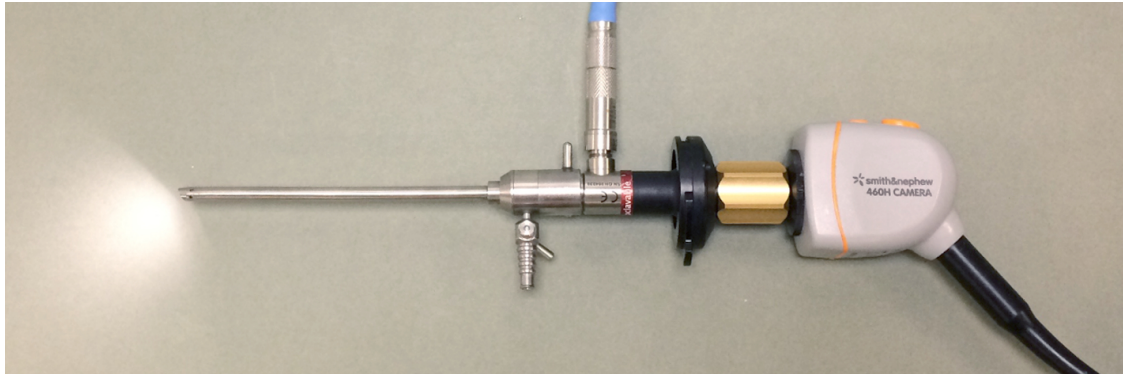
Metrics reported by WASP client v2.0.2	Metrics reported by Kirby et al ²²³
No. of hand movements; No. of minor hand movements; Smoothness of movement; Time taken; Maximum hand acceleration; Stationary time; Inactivity time; Hand dominance	Time taken; No. of hand movements; Smoothness of movement; Overall maximum hand acceleration; Range of accelerometer values; Range of gyroscope values; Variance of resultant acceleration; Variance of resultant gyroscope values; Intensity of movement in the 0-5Hz frequency band; Intensity of movement in the 5-20Hz frequency band; Dominance; Stationary time

A.3 Ethical approval

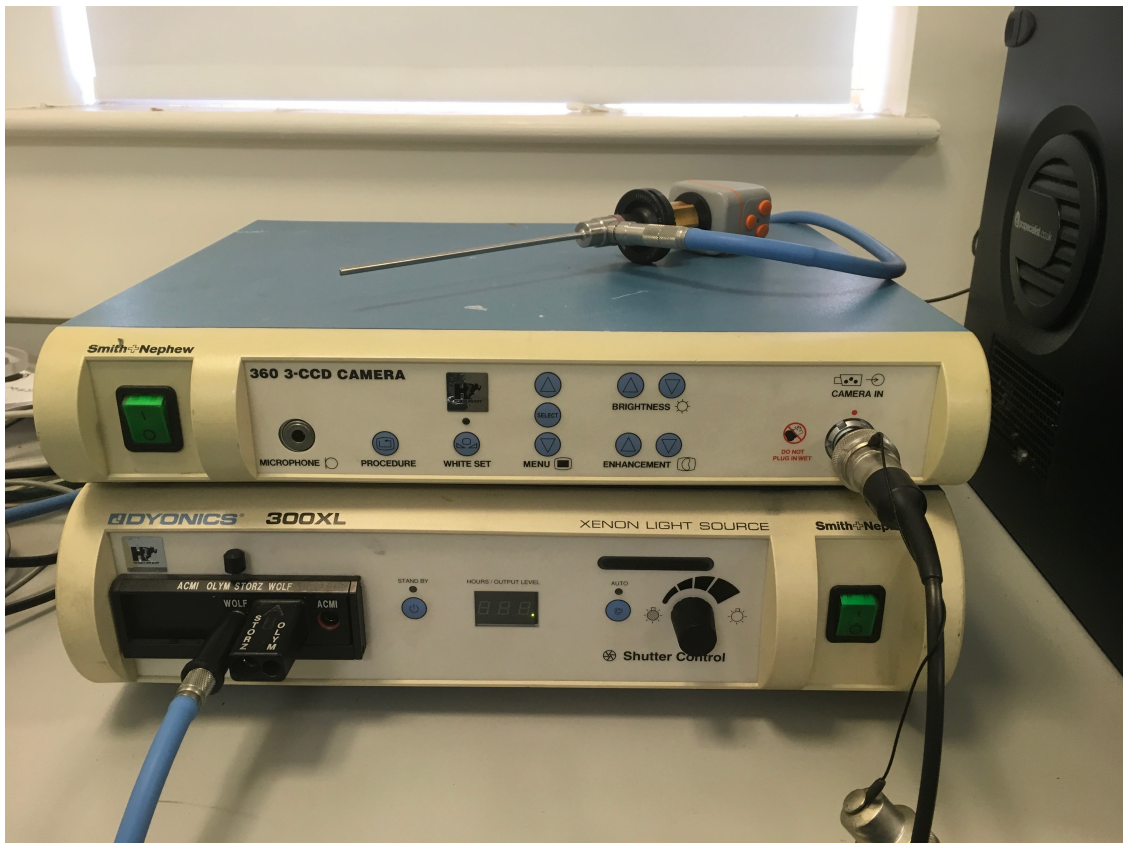
The studies within this thesis were conducting in accordance with ethical approval from the Central University Research Ethics Committee, a copy of which is reproduced in Figure A.4.

A.4 WASP motion analysis system

The WASP motion analysis system uses commercially available WAX9 sensors (Axivity, York, UK) which contain accelerometer, gyroscopic, magnetometer, temperature and pressure sensors (Figure A.5). These are attached to the participants elbow using silicone straps (Figure A.6). Accelerometer and gyroscope data are streamed over Bluetooth 4.0 BLE (low energy) to a computer running the custom WASP client (version 2.0.2; McLaren Applied Technologies, Woking, UK) which calculates motion analysis metrics from the raw streamed data (Table A.1). This data is displayed in the client so participants can see their most recent performance, their performance over time, and their performance in comparison to customisable benchmarks (Figure A.7).



(a) An arthroscope with camera and cannula



(b) Image management system and fibre-optic light source

Figure A.2: Arthroscopy equipment. (a) 30° arthroscope, cannula and camera. (b) Image management system and fibre-optic light source

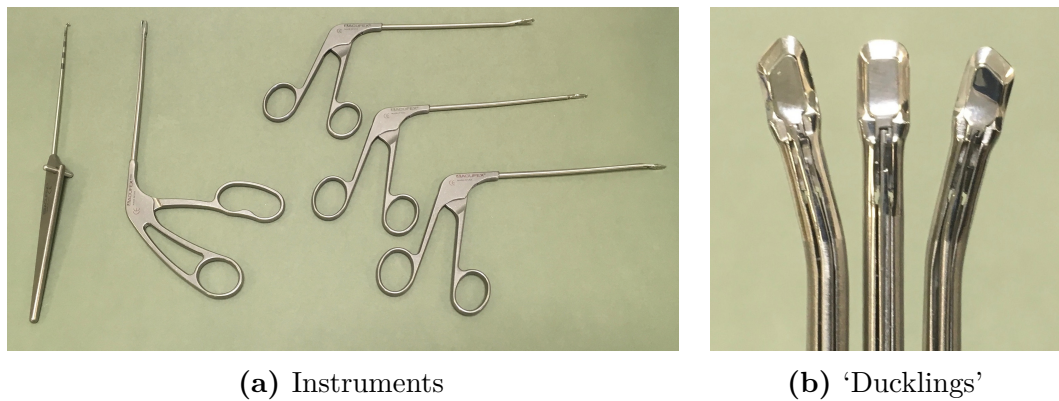


Figure A.3: Arthroscopy instruments. (a) From left; probe, grasper, and three 'duckling' meniscal punches (b) Detail of the three different angles of 'duckling' meniscal punches which allow participants to control the direction of debridement.

MEDICAL SCIENCES INTER DIVISIONAL RESEARCH ETHICS COMMITTEE
Research Services, University of Oxford

University Offices, Wellington Square, Oxford, OX1 2JD
Tel: +44(0)1865 616577 Fax: +44(0)1865 280467
ethics@medsci.ox.ac.uk <http://www.medsci.ox.ac.uk>



CONFIDENTIAL

Ref: MSD-IDREC-C1-2014-152

Professor Jonathan Rees
NDORMS
Botnar Institute of Musculoskeletal Sciences
University of Oxford
Old Road
Oxford

12th September 2014

Dear Professor Rees

CUREC checklist

I am writing to acknowledge receipt of your CUREC-1 form for your project: **Simulation based arthroscopic surgery**

On the basis of the information you have provided this has now been approved by the Medical Sciences IDREC **subject to:**

- a) **it is your responsibility to comply with the requirements for administering any tests or questionnaires and if in doubt to contact the publisher of those tests or questionnaires.**

The reference number for this project is **MSD-IDREC-C1-2014-152** and is valid for a period of **60 months** from the CUREC 1 approval date, **10th September 2014**. Please may I remind you that your project may be reviewed at some stage during an annual audit of projects.

Amendments

Should you at some stage alter some of the techniques or procedures then you should first undertake a checklist (CUREC-1) to see whether these changes alter the ethics of the research. If these remain the same then the committee will require notification of the changes to lodge with the project. If they do not remain the same then you may need to complete a CUREC-2 form and undergo further scrutiny by the committee.

Please do not hesitate to contact me if you have any queries about this.

Yours sincerely

A handwritten signature in blue ink, appearing to read 'Gill Halstead'.

Gill Halstead
Research Ethics Co-ordinator, Medical Sciences

Figure A.4: Letter granting ethical approval for the studies by the Central University Research Ethics Committee, reference number MSD-IDREC-C1-2014-152

**Specification: WAX9**

PARAMETER	VALUE	NOTES
Puck Size	23x32.5x8.9 mm	
Puck Weight	8g	
Enclosure Material	Polycarbonate	
Battery Capacity	85mAh	Rechargeable lithium polymer
Battery Charge Current	100mA	
Connectivity	Micro USB	
MAXIMUM LIFE (full charge)		
Hibernate	56 days	
Discoverable	117 hours	
Connected (not showing)	3 hours	Bluetooth Basic Rate
Connected (50Hz stream)	2 hours	Bluetooth Basic Rate
LE Mode Connected (not showing)	25 hours	Bluetooth 4.0 Low Energy
LE Mode Connected (50Hz stream)	6 hours	Bluetooth 4.0 Low Energy
ENVIRONMENTAL PROTECTION		
Dust Ingress	Dust tight (IP6X)	
Operating Temperature	0°C to 65°C (not charging)	40°C if charging
MEASUREMENT CAPABILITIES		
Output Data Rate	Configurable 1Hz - 400Hz	See developer guide for details
ACCELEROMETER		
Sensor Type	MMA8451Q	
Range	±2/4/8g	
Resolution	14-bit	up to 0.1mg
MAGNETOMETER		
Sensor Type	MAG 3110	
Range	±1mT	
Resolution	16-Bit	
GYRO		
Sensor Type	L3G4200D	
Range	±250/500/2000 dps	
Resolution	16-Bit	
TEMPERATURE		
Sensor Type	BMP180	Linear thermistor
Range	0°C - 65°C	
Resolution	0.1°C	
PRESSURE		
Sensor Type	BMP180	
Range	30-110 kPa	
Resolution	1Pa	

Figure A.5: Specifications of the WAX9 motion sensors

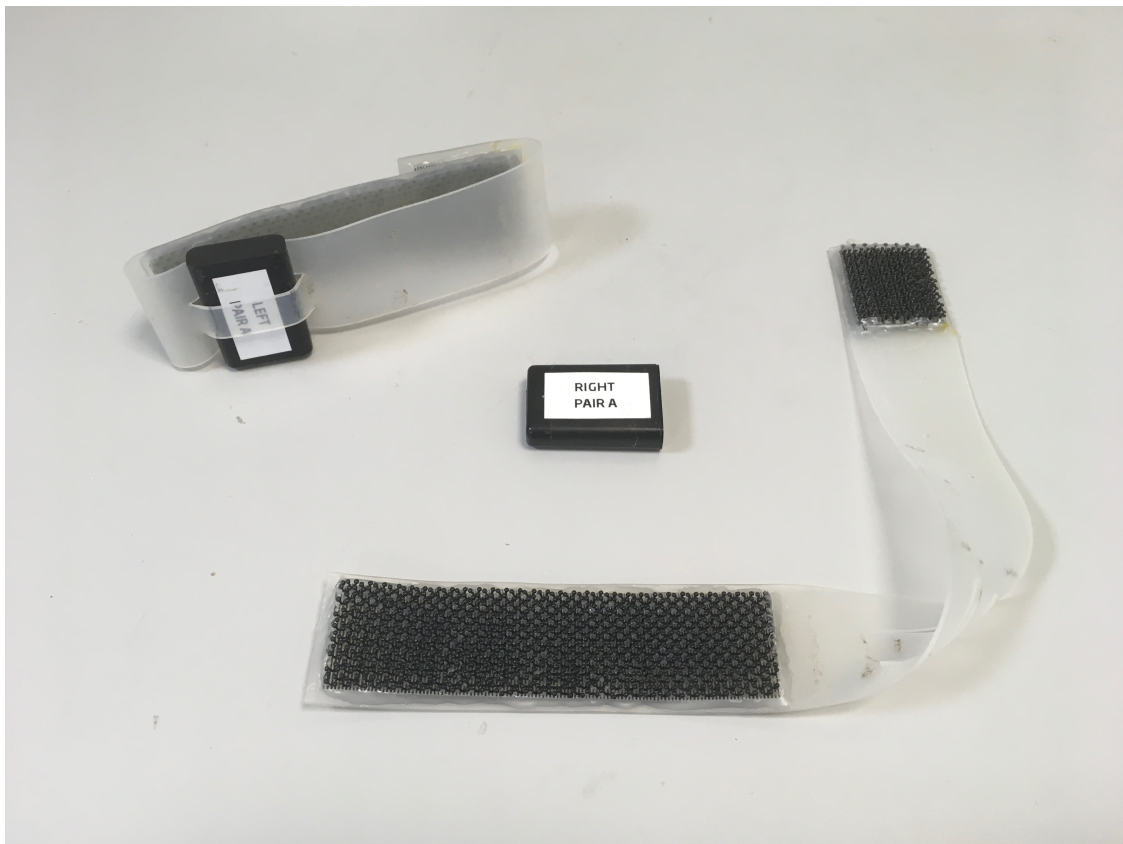


Figure A.6: WAX9 sensor positioning

WAX9 motion sensors are held in position at the participants elbows approx. three finger breadths from the lateral epicondyle using silicone straps

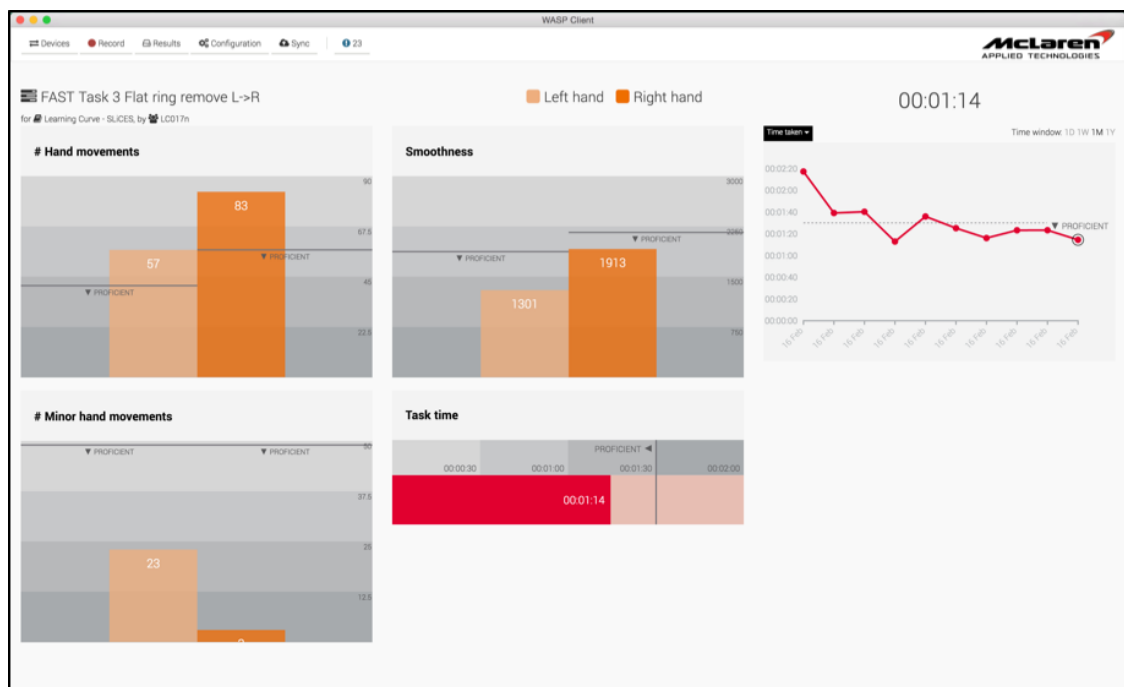


Figure A.7: The WASP client ‘dashboard’ allows participants to view their performance

B

Systematic review of quantitative methods of assessing arthroscopic performance

Contents

B.1	Search strategy	196
B.2	Full tables of study characteristics	197

This appendix contains details of the search strategy used to perform the systematic review in Chapter 2, and full data extracted for each of the included studies as a complement to the tables in Chapter 2 which contain the features of the studies as they apply to the specific quantitative metrics.

B.1 Search strategy

Medline via Ovid 1946 to week 4 of July 2016

1. exp Arthroscopy/ 19299
2. Arthroscopy.mp. 23556
3. exp Clinical competence/ 74962
4. exp Learning curve/ 1753
5. exp "Task Performance and Analysis"/ 30578
6. exp Competency-based Education/ 3201
7. exp Educational measurement/ 122181
8. exp Computer Simulation/184139
9. 1 or 2 23601
10. 3 or 4 or 5 or 6 or 7 or 8 334239
11. 9 and 10 259
12. limit 11 to English language 247

Embase via Ovid: 1946 to week 4 of July 2016

1. exp arthroscopy/ 23733
2. arthroscopy.mp. 28588
3. exp clinical competence/ 48864
4. exp task performance/ 124498
5. exp learning curve/ 6883
6. exp surgical training/ 15080
7. exp simulation/ 205447
8. 1 or 2 28588
9. 3 or 4 or 5 or 6 or 7 392522
10. 8 AND 9
11. limit 10 to English language 361

Medline via Pubmed: 1946 to week 4 of July 2016

1. exp Arthroscopy/ 18734
2. Arthroscopy.mp. 25353
3. exp Clinical competence/ 73768
4. exp Learning curve/ 1703
5. exp "Task Performance and Analysis"/ 30106
6. exp Competency-based Education/ 3141
7. exp Educational measurement/ 120485
8. exp Computer Simulation/180247
9. 1 or 2 23601
10. 3 or 4 or 5 or 6 or 7 or 8 328231
11. 9 and 10 251
12. limit 11 to English language 239

Figure B.1: Systematic review search strategy for Medline and Embase databases via Ovid, and Medline via PubMed

B.2 Full tables of study characteristics

Table B.1: Characteristics of studies where time is used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Akhtar 2013 ¹⁹²	Cohort	3	Validation	Knee	Sim	VR	37	3
Akhtar 2013 ¹⁹³	Cohort	3	Efficacy	Knee	Sim	VR	16	2
Akhtar 2016 ¹⁸⁹	RCT	2	Efficacy	Knee, Lap	Sim	VR	72	4
Alvand 2011 ¹¹⁰	RCT	2	Efficacy	Knee, Shoulder	Sim	BT	33	2
Alvand 2011 ¹⁰⁹	Cohort	3	Efficacy	Knee, Shoulder	Sim	BT	20	1
Alvand 2012 ¹⁹⁴	Cohort	3	Validation	Knee	Sim	BT	27	3
Alvand 2013 ¹⁹⁵	Cohort	3	Validation	Knee	Sim	BT	19	1
Anakwenze 2013 ¹⁹⁶	Case series	4	Efficacy	Shoulder	OR	OR	1	0
Andersen 2011 ¹⁹⁷	RCT	2	Efficacy	Shoulder	Sim	VR	21	3
Angelo 2015 ¹⁹⁸	Cohort	3	Validation	Shoulder	Sim	BT	19	2
Angelo 2015 ¹²⁹	RCT	2	Efficacy	Shoulder	Sim	BT/Cad	44	3
Angelo 2015 ¹⁹⁹	Cohort	3	Validation	Shoulder	Sim	Cad	22	2
Bayona 2008 ²⁰⁰	Cohort	3	Validation	Shoulder	Sim	VR	32	2
Bliss 2005 ²⁰¹	Cohort	3	Validation	Knee	Sim	VR	10	1
Braman 2015 ²⁰²	Cohort	3	Validation	Generic	Sim	BT	16	2
Butler 2013 ²⁰³	RCT	2	Efficacy	Knee	Sim	BT/Cad	14	2
Camp 2016 ²⁰⁴	RCT	2	Efficacy	Knee	Sim	VR/Cad	45	3
Cannon 2014 ¹⁴⁵	RCT	2	Efficacy	Knee	Sim/OR	VR/OR	48	2
Cannon 2014 ²⁰⁵	Cohort	3	Validation	Knee	Sim	VR	18	3
Castricini 2013 ²⁰⁶	Case series	4	Efficacy	Shoulder	OR	OR	1	0
Chang 2016 ²⁰⁷	Cohort	3	Validation	Knee	Sim	VR/BT	19	2
Cowan 2016 ⁵¹	Cohort	3	Other	Knee	Sim	VR	25	2
Cunningham 2016 ²⁰⁸	Case series	4	Efficacy	Shoulder	OR	OR	1	0
Del Piñal 2012 ²⁰⁹	Case series	4	Efficacy	Wrist	OR	OR	1	0
Dunn 2015 ¹⁸⁶	RCT	2	Efficacy	Shoulder	Sim/OR	VR/OR	17	2
Elliot 2012 ²¹⁰	Cohort	3	Validation	Knee	Sim	Cad	30	3
Escoto 2012 ²¹¹	Cohort	3	Validation	Knee	Sim	BT	15	2
Escoto 2013 ²¹²	Cohort	3	Validation	Knee	Sim	BT	13	1
Fucentese 2015 ²¹³	Cohort	3	Validation	Knee	Sim	VR	68	3
Gandhi 2015 ²¹⁴	Cohort	3	Validation	Shoulder	Sim	VR/PC	49	1
Gandhi 2015 ²¹⁵	Cohort	3	Efficacy	Shoulder	Sim	VR	65	1
Gilmer 2015 ²¹⁶	Cohort	3	Efficacy	Generic	Sim	BT	64	1
Gomoll 2007 ²¹⁷	Cohort	3	Validation	Shoulder	Sim	VR	43	4
Gomoll 2008 ²¹⁸	Cohort	3	Efficacy	Shoulder	Sim	VR	10	1
Guttman 2005 ¹²⁶	Case series	4	Efficacy	Shoulder	OR	OR	1	0
Henn 2013 ²¹⁹	RCT	2	Efficacy	Shoulder	Sim	VR/Cad	17	2
Howells 2008 ²²⁰	Cohort	3	Validation	Shoulder	Sim	BT	35	4
Howells 2008 ¹⁴⁴	RCT	2	Efficacy	Knee	Sim/OR	BT/OR	20	2
Howells 2009 ²²¹	Cohort	3	Validation	Shoulder	Sim	BT	6	1
Jacobsen 2015 ⁹¹	Cohort	3	Validation	Knee	Sim	VR	26	2
Kim 2010 ²²²	Case series	4	Efficacy	Elbow	OR	OR	1	0
Kirby 2015 ²²³	Cohort	3	Validation	Shoulder	Sim	VR	30	3

BT: Bench-top; Cad: Cadaveric; Lap: Laparoscopic; OR: Operating room; Sim: Simulation; VR: Virtual reality. Continued on next page.

Table B.1: Characteristics of studies where time is used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Koebler 2015 ²²⁴	Cohort	3	Validation	Knee, Shoulder	OR	OR	12	4
Konan 2011 ²²⁵	Case series	4	Efficacy	Hip	OR	OR	1	0
Lee 2013 ²²⁶	Case series	4	Efficacy	Hip	OR	OR	1	0
Martin 2011 ²²⁷	Cohort	3	Efficacy	Shoulder	Sim	VR/Cad	26	2
Martin 2012 ²²⁸	Cohort	3	Efficacy	Shoulder	Sim	VR	27	1
Martin 2015 ²²⁹	RCT	2	Efficacy	Ankle	Sim	BT/Cad	29	2
McCarthy 2006 ²³⁰	Cohort	3	Validation	Knee	Sim	VR	25	2
Middleton 2016 ²³¹	Cohort	3	Efficacy	Knee, Shoulder	Sim	VR	63	3
Nicandri 2015 ²³²	Cohort	3	Efficacy	Knee	Sim	VR/Cad	46	1
Pedowitz 2002 ²³³	Cohort	3	Validation	Shoulder	Sim	VR	78	3
Pollard 2012 ²³⁴	RCT	2	Efficacy	Hip	Sim	BT	20	2
Ponce 2014 ¹⁸⁴	Case series	4	Other	Shoulder	OR	OR	15	1
Price 2015 ⁸⁴	Cohort	3	Efficacy	Knee	Sim	BT	54	5
Rahm 2016 ²³⁵	Cohort	3	Efficacy	Knee	Sim	VR	20	1
Rebolledo 2015 ²³⁶	RCT	2	Efficacy	Knee, Shoulder	Sim	VR/Cad	14	2
Rose 2015 ²³⁷	Cohort	3	Validation	Generic	Sim	VR	30	3
Safir 2012 ¹⁸⁸	Cohort	3	Efficacy	Knee and Lap	Sim	VR/BT	0	2
Smith 1999 ²³⁸	Cohort	3	Validation	Shoulder	Sim	VR	18	3
Srivastava 2004 ²³⁹	Cohort	3	Validation	Shoulder	Sim	VR	35	3
Strom 2004 ¹⁸⁷	RCT	2	Efficacy	Multiple	Sim	VR	25	2
Stunt 2014 ²⁴⁰	Cohort	3	Validation	Knee	Sim	VR	27	3
Stunt 2014 ²⁴¹	Cohort	3	Validation	Knee	Sim	BT	31	3
Tashiro 2009 ²⁴²	Cohort	3	Validation	Knee	Sim	BT	30	3
Thomsen 2004 ¹⁸⁰	Cohort	3	Other	Spine	Sim	BT	15	1
Tuijthof 2010 ²⁴³	Cohort	3	Validation	Knee	Sim	BT	28	2
Tuijthof 2011 ²⁴⁴	Cohort	3	Validation	Knee	Sim	VR	37	3
Wang 2011 ²⁴⁵	Case series	4	Efficacy	Spine	OR	OR	1	0
Waterman 2016 ¹⁸⁵	RCT	2	Efficacy	Shoulder	Sim/OR	VR/OR	22	2
Wong 2015 ²⁴⁶	Cohort	3	Validation	Generic	Sim	BT	37	3
Yamakado 2015 ¹⁷⁹	Case series	4	Efficacy	Shoulder	OR	OR	1	0

BT: Bench-top; Cad: Cadaveric; Lap: Laparoscopic; OR: Operating room; Sim: Simulation; VR: Virtual reality.

Table B.2: Characteristics of studies where motion parameters are used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Akhtar 2013 ¹⁹²	Cohort	3	Validation	Knee	Sim	VR	37	3
Akhtar 2016 ¹⁸⁹	RCT	2	Efficacy	Knee, Lap	Sim	VR	72	4
Akhtar 2013 ¹⁹³	Cohort	3	Efficacy	Knee	Sim	VR	16	2
Alvand 2011 ¹¹⁰	RCT	2	Efficacy	Knee, Shoulder	Sim	BT	33	2
Alvand 2011 ¹⁰⁹	Cohort	3	Efficacy	Knee, Shoulder	Sim	BT	20	1
Alvand 2013 ¹⁹⁵	Cohort	3	Validation	Knee	Sim	BT	19	1
Alvand 2012 ¹⁹⁴	Cohort	3	Validation	Knee	Sim	BT	27	3
Andersen 2011 ¹⁹⁷	RCT	2	Efficacy	Shoulder	Sim	VR	21	3
Chang 2016 ²⁰⁷	Cohort	3	Validation	Knee	Sim	VR/BT	19	2
Escoto 2013 ²¹²	Cohort	3	Validation	Knee	Sim	BT	13	1
Fucentese 2015 ²¹³	Cohort	3	Validation	Knee	Sim	VR	68	3
Gandhi 2015 ²¹⁴	Cohort	3	Validation	Shoulder	Sim	VR/PC	49	1
Gandhi 2015 ²¹⁵	Cohort	3	Efficacy	Shoulder	Sim	VR	65	1
Gomoll 2007 ²¹⁷	Cohort	3	Validation	Shoulder	Sim	VR	43	4
Gomoll 2008 ²¹⁸	Cohort	3	Efficacy	Shoulder	Sim	VR	10	1
Howells 2008 ²²⁰	Cohort	3	Validation	Shoulder	Sim	BT	35	4
Howells 2009 ²²¹	Cohort	3	Validation	Shoulder	Sim	BT	6	1
Howells 2008 ¹⁴⁴	RCT	2	Efficacy	Knee	Sim/OR	BT/OR	20	2
Jackson 2012 ²⁴⁷	RCT	2	Efficacy	Knee	Sim	BT	19	3
Jacobsen 2015 ⁹¹	Cohort	3	Validation	Knee	Sim	VR	26	2
Kirby 2015 ²²³	Cohort	3	Validation	Shoulder	Sim	VR	30	3
Martin 2012 ²²⁸	Cohort	3	Efficacy	Shoulder	Sim	VR	27	1
McCarthy 2006 ²³⁰	Cohort	3	Validation	Knee	Sim	VR	25	2
Middleton 2016 ²³¹	Cohort	3	Efficacy	Knee, Shoulder	Sim	VR	63	3
Nicandri 2015 ²³²	Cohort	3	Efficacy	Knee	Sim	VR/Cad	46	1
Pedowitz 2002 ²³³	Cohort	3	Validation	Shoulder	Sim	VR	78	3
Pollard 2012 ²³⁴	RCT	2	Efficacy	Hip	Sim	BT	20	2
Price 2015 ⁸⁴	Cohort	3	Efficacy	Knee	Sim	BT	54	5
Rahm 2016 ²³⁵	Cohort	3	Efficacy	Knee	Sim	VR	20	1
Rose 2015 ²³⁷	Cohort	3	Validation	Generic	Sim	VR	30	3
Smith 1999 ²³⁸	Cohort	3	Validation	Shoulder	Sim	VR	18	3
Srivastava 2004 ²³⁹	Cohort	3	Validation	Shoulder	Sim	VR	35	3
Strom 2004 ¹⁸⁷	RCT	2	Efficacy	Multiple	Sim	VR	25	2
Tashiro 2009 ²⁴²	Cohort	3	Validation	Knee	Sim	BT	30	3
Waterman 2016 ¹⁸⁵	RCT	2	Efficacy	Shoulder	Sim/OR	VR/OR	22	2

BT: Bench-top; Cad: Cadaveric; Lap: Laparoscopic; OR: Operating room; Sim: Simulation; VR: Virtual reality.

Table B.3: Characteristics of studies where global rating scales are used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Alvand 2013 ¹⁹⁵	Cohort	3	Validation	Knee	Sim	BT	19	1
Alvand 2012 ¹⁹⁴	Cohort	3	Validation	Knee	Sim	BT	27	3
Bayona 2014 ²⁵⁵	Cohort	3	Validation	Shoulder	Sim	VR	49	3
Butler 2013 ²⁰³	RCT	2	Efficacy	Knee	Sim	BT/Cad	14	2
Camp 2016 ²⁰⁴	RCT	2	Efficacy	Knee	Sim	VR/Cad	45	3
Cannon 2014 ¹⁴⁵	RCT	2	Efficacy	Knee	Sim/OR	VR/OR	48	2
Chang 2016 ²⁰⁷	Cohort	3	Validation	Knee	Sim	VR/BT	19	2
Dunn 2015 ¹⁸⁶	RCT	2	Efficacy	Shoulder	Sim/OR	VR/OR	17	2
Dwyer 2015 ²⁵⁶	Cohort	3	Validation	Knee	Sim	BT	40	3
Dwyer 2016 ²⁵⁷	Cohort	4	Validation	Knee, Shoulder	Sim	BT	41	4
Henn 2013 ²¹⁹	RCT	2	Efficacy	Shoulder	Sim	VR/Cad	17	2
Hodgins 2014 ²⁵⁸	Cohort	3	Efficacy	Knee	OR	OR	20	1
Howells 2008 ¹⁴⁴	RCT	2	Efficacy	Knee	Sim/OR	BT/OR	20	2
Hoyle 2011 ²⁵¹	Cohort	3	Validation	Shoulder	OR	OR	10	1
Insel 2009 ²⁵⁰	Cohort	3	Validation	Knee	Sim	Cad	70	6
Kirby 2015 ²²³	Cohort	3	Validation	Shoulder	Sim	VR	30	3
Koehler 2013 ²⁴⁹	Cohort	3	Validation	Knee	Sim	Cad	30	3
Koehler 2015 ²²⁴	Cohort	3	Validation	Knee, Shoulder	OR	Theatre	12	4
Koehler 2013 ²⁵⁹	Cohort	3	Efficacy	Knee	Sim	Cad	30	1
Martin 2015 ²²⁹	RCT	2	Efficacy	Ankle	Sim	BT/Cad	29	2
Martin 2015 ¹³⁹	Cohort	3	Validation	Knee	Sim	Animal/Cad	15	4
Middleton 2016 ²³¹	Cohort	3	Efficacy	Knee, Shoulder	Sim	VR	63	3
Nicandri 2015 ²³²	Cohort	3	Efficacy	Knee	Sim	VR/Cad	46	1
Olson 2013 ²⁶⁰	Cohort	3	Efficacy	Knee	Sim	Cad	23	2
Price 2015 ⁸⁴	Cohort	3	Efficacy	Knee	Sim	BT	54	5
Slade Shantz 2013 ²⁵²	Cohort	3	Validation	Knee	Sim	Cad	20	4
Waterman 2016 ¹⁸⁵	RCT	2	Efficacy	Shoulder	Sim/OR	VR/OR	22	2
Wong 2013 ²⁴⁶	Cohort	3	Validation	Generic	Sim	BT	35	1

BT: Bench-top; Cad: Cadaveric; Lap: Laparoscopic; OR: Operating room; Sim: Simulation; VR: Virtual reality.

Table B.4: Characteristics of studies where task-specific checklists are used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Angelo 2015 ¹⁹⁸	Cohort	3	Validation	Shoulder	Sim	BT	19	2
Angelo 2015 ¹²⁹	RCT	2	Efficacy	Shoulder	Sim	BT/Cad	44	3
Angelo 2015 ¹⁹⁹	Cohort	3	Validation	Shoulder	Sim	Cad	22	2
Angelo 2015 ¹²⁸	Case series	4	Validation	Shoulder	OR	Videos	14	1
Bliss 2005 ²⁰¹	Cohort	3	Validation	Knee	Sim	VR	10	1
Cannon 2014 ¹⁴⁵	RCT	2	Efficacy	Knee	Sim/OR	VR/OR	48	2
Cowan 2016 ⁵¹	Cohort	3	Other	Knee	Sim	VR	25	2
Dunn 2015 ¹⁸⁶	RCT	2	Efficacy	Shoulder	Sim/OR	VR/OR	17	2
Dwyer 2015 ²⁵⁶	Cohort	3	Validation	Knee	Sim	BT	40	3
Dwyer 2016 ²⁵⁷	Cohort	4	Validation	Knee, Shoulder	Sim	BT	41	4
Elliot 2012 ²¹⁰	Cohort	3	Validation	Knee	Sim	Cad	30	3
Hodgins 2014 ²⁵⁸	Cohort	3	Efficacy	Knee	OR	OR	20	1
Howells 2008 ¹⁴⁴	RCT	2	Efficacy	Knee	Sim/OR	BT/OR	20	2
Insel 2009 ²⁵⁰	Cohort	3	Validation	Knee	Sim	Cad	70	6
Koehler 2013 ²⁴⁹	Cohort	3	Validation	Knee	Sim	Cad	30	3
Martin 2015 ²²⁹	RCT	2	Efficacy	Ankle	Sim	BT/Cad	29	2
Martin 2015 ¹³⁹	Cohort	3	Validation	Knee	Sim	Animal/Cad	15	4
McCarthy 2006 ²³⁰	Cohort	3	Validation	Knee	Sim	VR	25	2
Slade Shantz 2013 ²⁵²	Cohort	3	Validation	Knee	Sim	Cad	20	4
Smith 1999 ²³⁸	Cohort	3	Validation	Shoulder	Sim	VR	18	3
Srivastava 2004 ²³⁹	Cohort	3	Validation	Shoulder	Sim	VR	35	3
Talbot 2015 ²⁶³	Cohort	3	Validation	Shoulder	OR	OR	6	3
Thomsen 2004 ¹⁸⁰	Cohort	3	Other	Spine	Sim	BT	15	1
Waterman 2016 ¹⁸⁵	RCT	2	Efficacy	Shoulder	Sim/OR	VR/OR	22	2
Wong 2013 ²⁴⁶	Cohort	3	Validation	Generic	Sim	BT	35	1

BT: Bench-top; Cad: Cadaveric; Lap: Laparoscopic; OR: Operating room; Sim: Simulation; VR: Virtual reality.

Table B.5: Characteristics of studies where error lists are used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Angelo 2015 ¹⁹⁸	Cohort	3	Validation	Shoulder	Sim	BT	19	2
Angelo 2015 ¹²⁹	RCT	2	Efficacy	Shoulder	Sim	BT Cad	44	3
Angelo 2015 ¹⁹⁹	Cohort	3	Validation	Shoulder	Sim	Cad	22	2
Angelo 2015 ¹²⁸	Case series	4	Validation	Shoulder	OR	Videos	14	1

BT: Bench-top; Cad: Cadaveric; OR: Operating room; Sim: Simulation.

Table B.6: Characteristics of studies where behavioural features are used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Alvand 2012 ¹⁹⁴	Cohort	3	Validation	Knee	Sim	BT	27	3
Andersen 2011 ¹⁹⁷	RCT	2	Efficacy	Shoulder	Sim	VR	21	3
Bliss 2005 ²⁰¹	Cohort	3	Validation	Knee	Sim	VR	10	1
Braman 2015 ²⁰²	Cohort	3	Validation	Generic	Sim	BT	16	2
Chang 2016 ²⁰⁷	Cohort	3	Validation	Knee	Sim	VR/BT	19	2
Cowan 2016 ⁵¹	Cohort	3	Other	Knee	Sim	VR	25	2
Fucentese 2015 ²¹³	Cohort	3	Validation	Knee	Sim	VR	68	3
Gandhi 2015 ²¹⁴	Cohort	3	Validation	Shoulder	Sim	VR/PC	49	1
Gomoll 2007 ²¹⁷	Cohort	3	Validation	Shoulder	Sim	VR	43	4
Gomoll 2008 ²¹⁸	Cohort	3	Efficacy	Shoulder	Sim	VR	10	1
McCarthy 2006 ²³⁰	Cohort	3	Validation	Knee	Sim	VR	25	2
Pedowitz 2002 ²³³	Cohort	3	Validation	Shoulder	Sim	VR	78	3
Smith 1999 ²³⁸	Cohort	3	Validation	Shoulder	Sim	VR	18	3
Srivastava 2004 ²³⁹	Cohort	3	Validation	Shoulder	Sim	VR	35	3
Strom 2004 ¹⁸⁷	RCT	2	Efficacy	Multiple	Sim	VR	25	2

BT: Bench-top; Sim: Simulation; VR: Virtual reality.

Table B.7: Characteristics of studies where force parameters are used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Andersen 2011 ¹⁹⁷	RCT	2	Efficacy	Shoulder	Sim	VR	21	3
Chami 2008 ⁷¹	Cohort	3	Validation	Knee	OR	OR	5	2
Escoto 2012 ²¹¹	Cohort	3	Validation	Knee	Sim	BT	15	2
Escoto 2013 ²¹²	Cohort	3	Validation	Knee	Sim	BT	68	3
Gilmer 2015 ²¹⁶	Cohort	3	Efficacy	Generic	Sim	BT	64	1
Nicandri 2015 ²³²	Cohort	3	Efficacy	Knee	Sim	VR/Cad	46	1
Pedowitz 2002 ²³³	Cohort	3	Validation	Shoulder	Sim	VR	78	3
Pedowitz 2015 ²⁶⁵	RCT	3	Efficacy	Generic	Sim	BT	94	5
Smith 1999 ²³⁸	Cohort	3	Validation	Shoulder	Sim	VR	18	3
Srivastava 2004 ²³⁹	Cohort	3	Validation	Shoulder	Sim	VR	35	3
Tashiro 2009 ²⁴²	Cohort	3	Validation	Knee	Sim	BT	30	3

BT: Bench-top; Cad: Cadaveric; Lap: Laparoscopic; OR: Operating room; Sim: Simulation; VR: Virtual reality.

Table B.8: Characteristics of studies where single-item scales are used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Dunn 2015 ¹⁸⁶	RCT	2	Efficacy	Shoulder	Sim/OR	VR/OR	17	2
Rebolledo 2015 ²³⁶	RCT	2	Efficacy	Knee, Shoulder	Sim	VR/Cad	14	2
Slade Shantz 2013 ²⁵²	Cohort	3	Validation	Knee	Sim	Cad	20	4
Tashiro 2009 ²⁴²	Cohort	3	Validation	Knee	Sim	BT	30	3
Unalan 2010 ²⁶⁶	Cohort	3	Validation	Knee	Sim	Animal	64	1
Waterman 2016 ¹⁸⁵	RCT	2	Efficacy	Shoulder	Sim/OR	VR/OR	22	2

BT: Bench-top; Cad: Cadaveric; OR: Operating room; Sim: Simulation; VR: Virtual reality.

Table B.9: Characteristics of studies where VR proprietary metrics are used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Cannon 2014 ²⁰⁵	Cohort	3	Validation	Knee	Sim	VR	18	3
McCarthy 2006 ²³⁰	Cohort	3	Validation	Knee	Sim	VR	25	2
Nicandri 2015 ²³²	Cohort	3	Efficacy	Knee	Sim	VR/Cad	46	1
Rose 2015 ²³⁷	Cohort	3	Validation	Generic	Sim	VR	30	3
Srivastava 2004 ²³⁹	Cohort	3	Validation	Shoulder	Sim	VR	35	3
Strom 2004 ¹⁸⁷	RCT	2	Efficacy	Multiple	Sim	VR	25	2

BT: Bench-top; Cad: Cadaveric; OR: Operating room; Sim: Simulation; VR: Virtual reality.

Table B.10: Characteristics of studies where calculated scores are used as quantitative performance metrics

<i>Paper</i>	<i>Study type</i>	<i>OCEBM</i>	<i>Purpose</i>	<i>Joint</i>	<i>Setting</i>	<i>Environment</i>	<i>Participants</i>	<i>Groups</i>
Bayona 2008 ²⁰⁰	Cohort	3	Validation	Shoulder	Sim	VR	32	2
Cannon 2014 ²⁰⁵	Cohort	3	Validation	Knee	Sim	VR	18	3
Chang 2016 ²⁰⁷	Cohort	3	Validation	Knee	Sim	VR/BT	19	2
Coughlin 2015 ²⁶⁸	Cohort	3	Validation	Generic	Sim	BT	49	4
Elliot 2012 ²¹⁰	Cohort	3	Validation	Knee	Sim	Cad	30	3
Ferkel 2013 ²⁶⁹	Cohort	3	Validation	Shoulder	Sim	VR	41	1
Gandhi 2015 ²¹⁵	Cohort	3	Efficacy	Shoulder	Sim	VR	65	1
Jacobsen 2015 ⁹¹	Cohort	3	Validation	Knee	Sim	VR	26	2
Kirby 2015 ²²³	Cohort	3	Validation	Shoulder	Sim	VR	30	3
Rahm 2016 ²³⁵	Cohort	3	Efficacy	Knee	Sim	VR	20	1
Strom 2004 ¹⁸⁷	RCT	2	Efficacy	Multiple	Sim	VR	25	2
Wong 2013 ²⁴⁶	Cohort	3	Validation	Generic	Sim	BT	35	1

BT: Bench-top; Cad: Cadaveric; OR: Operating room; Sim: Simulation; VR: Virtual reality.

C

Development of a simulated arthroscopy list

Contents

C.1	FAST task selection survey	206
C.2	Meniscectomy task development	206
	C.2.1 Process and materials	206
	C.2.2 Meniscal tear morphology	208
C.3	Face and content validity survey	210
C.4	Construct validation results tables	212

C.1 FAST task selection survey

Question 1 stem How would you rate the complexity of the tasks you have performed?
(Multiple choice)

Question items Task 1 - Probe; Task 2 - Maze; Task 3 - Flat ring transfer; Task 4 - Column ring transfer; Task 5 - Arc cutting; Task 6 - Knot tying; Task 7 - Cuff repair

Response items Basic; Intermediate; Advanced

Question 2 stem Rank which modules would be most useful for novice trainees? e.g. Med students, Foundation & early Core trainees (Multiple choice)

Question items Task 1 - Probe; Task 2 - Maze; Task 3 - Flat ring transfer; Task 4 - Column ring transfer; Task 5 - Arc cutting; Task 6 - Knot tying; Task 7 - Cuff repair

Response items 1 - Most useful; 2; 3; 4; 5; 6; 7 - Least useful

Question 3 stem Which modules would you suggest to train the skills required for diagnostic and simple therapeutic knee arthroscopic procedures (e.g. loose body retrieval, meniscectomy) to core surgical trainees? (Checkbox)

Response items Task 1 - Probe; Task 2 - Maze; Task 3 - Flat ring transfer; Task 4 - Column ring transfer; Task 5 - Arc cutting; Task 6 - Knot tying; Task 7 - Cuff repair

Question 4 stem Do you feel Tasks 3 and 4 (the flat and column ring transfer modules) are sufficiently different, or develop separate skills, that you would include both in a training programme? (Multiple choice)

Response items Yes; No

C.2 Meniscectomy task development

Arthroscopic meniscectomy and partial meniscectomy are common procedures,⁵⁴ and all trainees are expected to reach the highest level of competency by the completion of training.⁸⁹ This study required a cost-effective method of producing a meniscectomy task at high volume for deliberate practice of this essential procedure.

C.2.1 Process and materials

Meniscal inserts are soft plastic, rubber, or silicone simulations of the medial and lateral menisci and cruciate ligaments for use with the #1401 Sawbones Knee simulator. They can be purchased with pre-formed meniscal tears, or intact. Each insert is only usable once, as the meniscectomy simulation removes parts of the insert irreparably. The studies in Chapters 4 and 6 were anticipated to require at

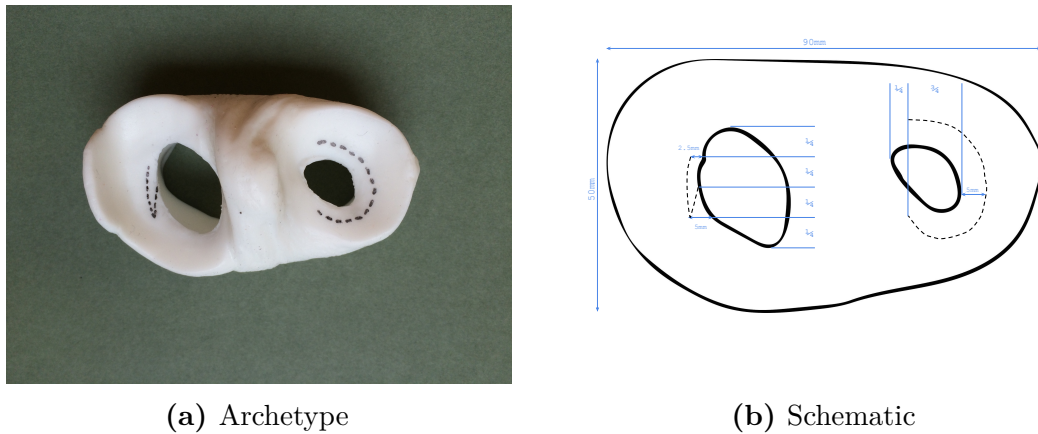


Figure C.1: Guides for creating reproducible bucket-handle and vertical oblique flap tears

least 100 inserts. To maintain a sustainable supply, three sources of meniscal insert were considered: commercial purchase, 3D-printing, or relief moulding.

Commercial purchase of meniscal inserts from Sawbones is the quickest way of acquiring large volumes but is also the most costly at 16USD per item (approx. 10GBP at the time of the study). 3D-printing requires a computer-assisted design file of the tissues intended for printing, and in the case of menisci these can be created with DICOM files from MRI scans. Although this gives the advantage of being able to simulate ‘real’ meniscal tears found in patients, the attempts at the actual 3D-printing process using an Ultimaker 2 Extended 3D-printer were both slow and temperamental, with prints taking many hours and prone to failure during extrusion when using the flexible filaments required to simulate the texture of meniscal fibrocartilage (Figure C.2). Relief moulding uses silicone rubber to make mould block of the inside and outside of a commercially purchased meniscal insert (Figure C.3). Once cured, these two mould blocks can be filled with more silicone rubber, which cures in the shape of the original insert. Six meniscal inserts could be reliably made in the working day, a process slower than commercial purchase but quicker than 3D-printing, and costing about 50 pence per insert, significantly cheaper than both. The process produced intact meniscal inserts which would need tears to be manually created for the simulations (Figure C.1).

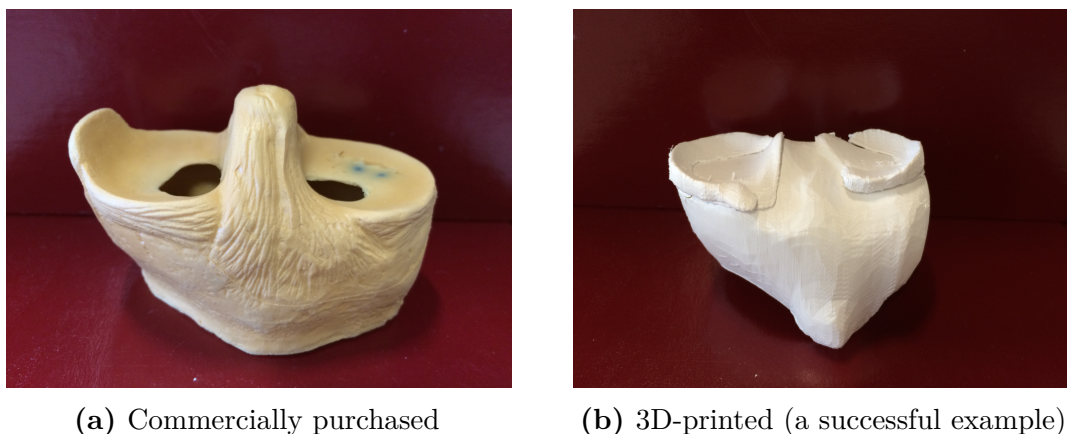


Figure C.2: Rejected sources of meniscal inserts

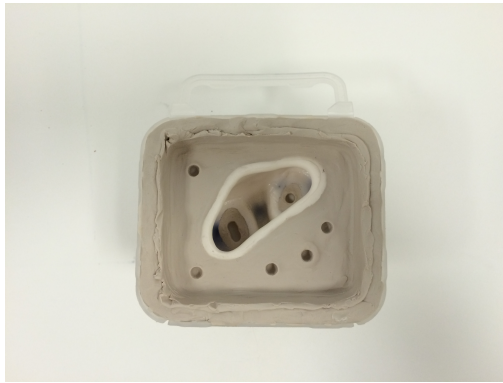
C.2.2 Meniscal tear morphology

This study used a displaced “bucket-handle” tear and a displaced vertical oblique flap as the simulated tear morphologies (Figure C.1). These were chosen from a common tear morphology classification system³³⁷ (Figure C.4) based on three factors: reproducibility, maximising technical skill requirement, and tears being appropriate for meniscectomy.

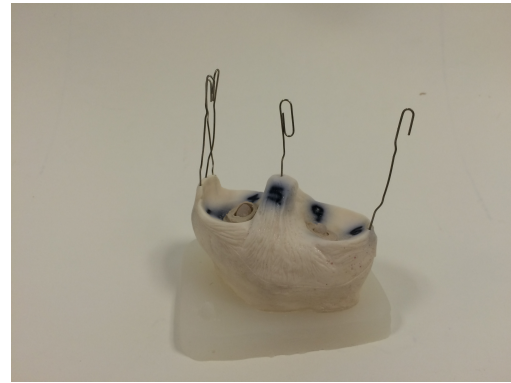
Reproducibility was important because the tears are manually created as the moulds produced intact the meniscal inserts. Tears needed to be easy to reproduce manually to ensure each participant has a fair opportunity at the same task. Horizontal tears, horizontal flaps and complex and degenerative tears were excluded as these are challenging to create consistently.

This task was designed as test of technical skill, so to increase the skill required to complete the task both tear morphologies were displaced: the bucket-handle tear interposed between the femoral condyle, and the apex of the vertical oblique flap displace under the intact meniscal rim. This required the additional step of reducing the tear to its anatomic location rather than just debriding the meniscus *in situ*, which favoured the bucket handle and vertical oblique morphologies, as radial tears do not typically displace.

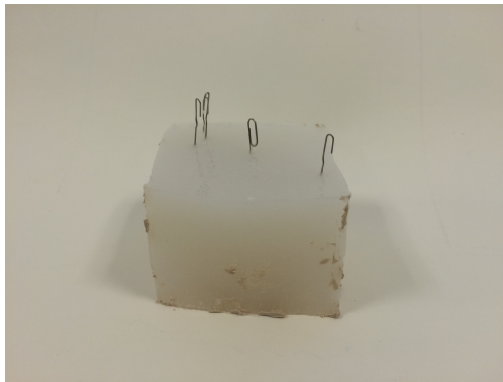
Finally, consideration was given to current controversy regarding management of meniscal tears,³³⁶ and the intention that this simulation be a purely technical



(a) Mould base. Plasticine is used to restrict flow of silicone



(b) Commercial meniscal insert on base. Outer mould to be cast around this.



(c) Outer mould casing has cured. Exposed paperclips create air channels



(d) Complete mould set

Figure C.3: Taking a relief mould of a commercial meniscal insert.

rather than decision-making exercise. Displaced tears cause mechanical symptoms and necessitate surgical intervention, further supporting the morphologies chosen, while tears were considered to be in the avascular (in “white zone”) and in older patients to rule out the chance of successful meniscal repair. For more experienced participants who may consider other treatment options, this was important to reduce the ambiguity of the task at hand. For inexperienced participants, when practicing technical skills in isolation it is important use appropriate scenarios so as not to inadvertently influence future decision-making behaviour or train bad habits, a valid concern around surgical simulation.¹⁵⁷

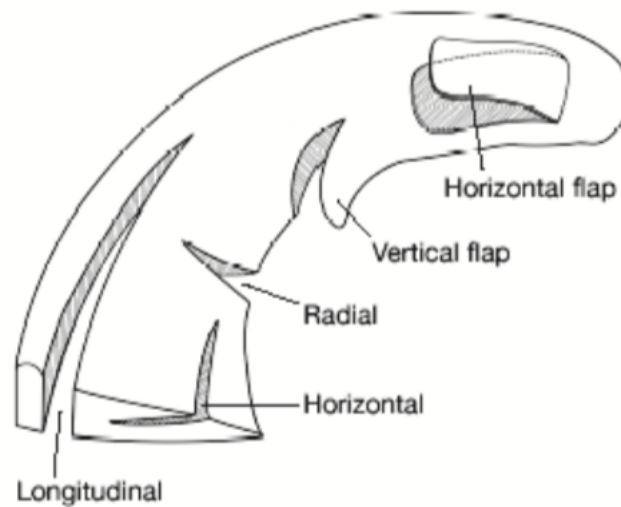


Figure C.4: Common meniscal tear morphologies. (Reproduced from Anderson et al³³⁷ with permission).

C.3 Face and content validity survey

Question 1 stem How do you feel about the way the results are presented regards improving your performance? (Multiple choice)

Response items Motivating; Ambivalent; Demotivating; Don't know/no opinion

Question 2 stem Having performed some arthroscopic tasks with motion analysis feedback, how useful do you anticipate these motion analysis parameters might be for training? (Multiple choice)

Question items Time taken; Numbers of hand movements; Smoothness of movements; Maximum acceleration; Stationary/Inactivity time; Difference in hand dominance

Response items Detrimental; Potentially detrimental; Neither useful nor detrimental; Potentially useful; Useful

Question 3 stem Is it clear how the metrics represent performance? (Multiple choice)

Question items Time taken; Numbers of hand movements; Smoothness of movements; Maximum acceleration; Stationary/Inactivity time; Difference in hand dominance

Response items Intuitive - this metric is obvious to me; Understandable - I can rationalise what this metric implies; Confusing - I don't know how to interpret this metric as a measure of performance; Irrelevant - I don't see how this is relevant to performance; Don't know/no opinion

Question 4 stem Is it clear how you could improve your performance? (Multiple choice)

Question items Time taken; Numbers of hand movements; Smoothness of movements; Maximum acceleration; Stationary/Inactivity time; Difference in hand dominance

Response items I know what I need to do to improve; I understand the metric, but don't know how I can improve on it; I can't interpret this metric, or how it might be improved; Don't know/no opinion

Question 5 stem What impact do you think time spent on this FAST simulator would have on these general arthroscopic skills? (Multiple choice)

Question items Anatomical knowledge; Triangulation; Spatial perception; Tissue manipulation; Manual dexterity

Response items Tend to impair; None or minimal impact; Tend to improve

Question 6 stem What impact do you think time spent on the knee meniscectomy simulator would have on these general arthroscopic skills? (Multiple choice)

Question items Anatomical knowledge; Triangulation; Spatial perception; Tissue manipulation; Manual dexterity

Response items Tend to impair; None or minimal impact; Tend to improve

Question 7 stem What impact do you think time spent on the knee meniscectomy simulator would have on these specific arthroscopic skills? (Multiple choice)

Question items Identification of structures and navigation with the arthroscope; Instrument handling; Preparation of patient and equipment; Portal placement; Inspection with the arthroscope; Navigation by visualisation of structures and probing them; Insertion of arthroscope into anterolateral portal; Meniscectomy; Meniscal suturing

Response items Tend to impair; None or minimal impact; Tend to improve

Question 8 stem What is your opinion of the realism of the following features of the simulation? (Multiple choice)

Question items Textures of the structures; Colours of the structures; Size of the structures; Size of the joint space; Arthroscopic display; Arthroscope itself; Instruments; 'Feel' when probing the meniscus; 'Feel' when cutting the meniscus

Response items Unrealistic, and detracts from the purpose of simulation; Unrealistic, but does not detract from the purpose of simulation; Realistic (or sufficient realism for the purposes of the simulation); Don't know/No opinion

Question 9 stem What is your opinion of the appearance of the meniscectomy simulation? (Multiple choice)

Question items Internal appearance; External appearance

Response items Obvious - it looks like a knee joint; Implied - some features that suggest it to be a knee joint; Vague - no features which suggest it's any particular joint; Ambiguous - mixed features suggesting it could be a knee, and one or more other joints; Deceptive - appearance of another joint

Question 10 stem Are all essential anatomic structures present? (Multiple choice & free text)

Response items Yes; Other... [Free text]

Question 11 stem Would you have used this simulator during your training if it had been available? (Multiple choice)

Response items Yes; No

Question 12 stem At what stages do you think this simulator is a suitable training modality in the skills lab? (Multiple choice)

Response items Novices - Medical Students, Foundation and Core Trainees; Intermediate - Registrars; Experts - Fellows and Consultants; Not appropriate for training

C.4 Construct validation results tables

Below are the results for each metric of the three groups performance on the simulated arthroscopy list tasks in the construct validation study in Chapter 4.

Table C.1: Performance by group, measured by time taken (s; median and IQR) on each of the simulated tasks. Where p-values are in parentheses they do not reach Bonferroni corrected significance (<0.0071).

Task	Time, s; median (IQR)			p-value		
	CCT	Intermediate	Novice	All-groups	Group-pair; CCT & Int	Group-pair; Int & Nov
Probe	249.4 (230.4-269.3)	311.7 (286.7-331.7)	608.0 (487.7-743.6)	0.000041	0.00071	0.000021
Maze	136.0 (104.7-156.7)	136.1 (104.6-159.7)	147.6 (132.6-166.3)	(0.52)		
Ring remove	90.1 (66.3-102.0)	83.7 (70.2-95.0)	197.3 (153.9-260.7)	0.00059	(0.95)	0.00012
Ring return	104.1 (94.0-144.3)	148.2 (118.0-234.2)	335.4 (252.9-396.1)	0.00070	(0.21)	0.0065
Arc Cutting	151.1 (116.6-180.4)	202.0 (156.4-224.9)	377.0 (269.5-479.6)	0.00015	(0.03)	0.0011
Diagnostic	67.4 (55.0-82.8)	111.3 (85.8-134.0)	357.3 (348.8-452.6)	0.000036	0.00074	0.000062
Menisectomy	205.5 (168.1-239.5)	329.9 (242.7-386.4)	928.7 (699.4-1042.7)	0.000012	0.0062	0.000011

Table C.2: Performance by group, measured by hand movements (n; median and IQR) on each of the simulated tasks. Where p-values are in parentheses they do not reach corrected significance (<0.0071).

Task	Hand movements, n; median (IQR)			p-value		
	CCT	Intermediate	Novice	All-groups	Group-pair; CCT & Int	Group-pair; Int & Nov
Probe	383 (346-399.5)	452 (440-514)	870 (761-1187)	0.000041	0.00071	0.000021
Maze	72.5 (52.25-107.25)	97 (63-136.75)	124.5 (107.25-143.75)	(0.041)		
Ring remove	93 (74.5-106)	114 (98-131)	249.5 (203-300.75)	0.00024	(0.12)	0.00047
Ring return	118 (100-146.5)	186 (107-248)	483.5 (349-574.75)	0.00031	(0.099)	0.007
Arc Cutting	133 (121-178)	218 (188-235)	398 (329-675.5)	0.000078	(0.068)	0.00034
Diagnostic	124 (100-137.5)	199 (149-220.75)	641.5 (604.75-751.75)	0.000005	0.0031	0.0000062
Menisectomy	349 (314.75-383.25)	577 (422.5-630.5)	1460 (1084.5-1881.25)	0.00001	(0.01)	0.000023

Table C.3: Performance by group, measured by minor hand movements (n; median and IQR) on each of the simulated tasks. Where p-values are in parentheses they do not reach corrected significance (<0.0071).

Task	Minor hand movements, n; median (IQR)			p-value		
	CCT	Intermediate	Novice	All-groups	Group-pair; CCT & Int	Group-pair; Int & Nov
Probe	135 (129-216.5)	221 (161-270)	453.5 (308.75-511)	0.00028	(0.072)	0.00086
Maze	204 (130.5-223.5)	177.5 (144-223.25)	180 (159-183.5)	(0.99)		
Ring remove	88 (72-102)	74 (60-83)	198.5 (124-246)	0.00062	(0.082)	0.00057
Ring return	102 (92.5-162)	138 (80-192)	231.5 (160.5-319.25)	(0.012)		
Arc Cutting	163 (128.5-208.5)	226 (213-244)	363.5 (291.25-480)	0.0027	(0.026)	(0.028)
Diagnostic	25 (21-30.5)	46.5 (41.5-73.25)	204 (160.75-247)	0.0000056	0.0014	0.00032
Meniscectomy	107 (77.75-160.75)	203 (143.5-275.5)	526.5 (391.5-691)	0.000013	(0.017)	0.0000057

Table C.4: Performance by group, measured by smoothness (ms^{-3} ; median and IQR) on each of the simulated tasks. Where p-values are in parentheses they do not reach corrected significance (<0.0071).

Task	Smoothness; ms^{-3} , median (IQR)			p-value	
	CCT	Intermediate	Novice	All- groups	Group-pair; CCT & Int Group-pair; Int & Nov
Probe	13093 (12068-14448)	18202 (14407-20003)	42841.5 (40073-51835)	0.00021	(0.072)
Maze	5088 (3903-5973)	5440.5 (4180-5818)	8151 (7196-8760)	0.0038	(0.54)
Ring remove	3344 (2894-4802)	3629 (3223-4722)	11749 (9694-16715)	0.00033	(0.95)
Ring return	5644 (3344-7754)	5744 (3678-12585)	24342 (18135-27891)	0.00045	(0.3)
Arc Cutting	7309 (5869-9056)	9330 (8333-11443)	24384 (16338.5-36618)	0.00037	(0.17)
Diagnostic	4801 (3161-5093)	6799 (5805-8697)	27920 (22772-37083)	0.000005	0.0028
Meniscec- tomy	16117 (12714-19297)	21844 (17639-26679)	65684 (51917-86956)	0.000024	(0.029)

Table C.5: Performance by group, measured by maximum acceleration (g; median and IQR) on each of the simulated tasks. Where p-values are in parentheses they do not reach corrected significance (<0.0071).

Task	Maximum acceleration;g; median (IQR)			p-value		
	CCT	Intermediate	Novice	All-groups	Group-pair; CCT & Int	Group-pair; Int & Nov
Probe	1.13 (1.10-1.18)	1.16 (1.08-1.22)	1.23 (1.19-1.35)	(0.03)		
Maze	1.12 (1.09-1.17)	1.15 (1.10-1.17)	1.19 (1.15-1.25)	(0.075)		
Ring remove	1.12 (1.08-1.14)	1.13 (1.10-1.17)	1.16 (1.13-1.20)	(0.097)		
Ring return	1.08 (1.07-1.12)	1.12 (1.08-1.15)	1.30 (1.22-1.37)	0.00045	(0.23)	0.0014
Arc Cutting	1.16 (1.12-1.26)	1.18 (1.11-1.24)	1.22 (1.20-1.34)	(0.12)		
Diagnostic	1.11 (1.09-1.20)	1.15 (1.13-1.20)	1.31 (1.22-1.42)	0.00088	(0.54)	0.0011
Meniscectomy	1.27 (1.25-1.33)	1.23 (1.20-1.28)	1.36 (1.33-1.41)	(0.028)		

Table C.6: Performance by group, measured by stationary time (% total time; median and IQR) on each of the simulated tasks. Where p-values are in parentheses they do not reach corrected significance (<0.0071).

Task	Stationary time;% total time, median (IQR)			p-value
	CCT	Intermediate	Novice	
Probe	4.6 (3.1-6.1)	6.4 (4.9-8.3)	7.2 (4.9-8.8)	(0.22)
Maze	16.7 (12.1-26.5)	21.8 (17.0-29.3)	20.1 (10.3-26.2)	(0.69)
Ring remove	10.7 (7.3-11.7)	5.3 (3.5-9.6)	8.9 (7.7-11.3)	(0.16)
Ring return	7.9 (7.3-10.2)	9.0 (5.1-11.5)	6.4 (4.3-10.5)	(0.48)
Arc Cutting	18.8 (10.5-24.6)	17.7 (12.6-33.7)	14.2 (10.9-18.2)	(0.42)
Diagnostic	1.1 (0.7-3.3)	2.5 (1.9-3.5)	4.0 (2.8-5.1)	(0.063)
Menisectomy	3.7 (3.2-5.3)	5.7 (3.2-6.3)	4.5 (3.7-6.6)	(0.48)

Table C.7: Performance by group, measured by inactivity time (% total time; median and IQR) on each of the simulated tasks. Where p-values are in parentheses they do not reach corrected significance (<0.0071).

Task	Inactivity time;% total time, median (IQR)			p-value	
	CCT	Intermediate	Novice	All-groups	CCT & Int Int & Nov
Probe	0.03 (0.005-0.06)	0.08 (0.02-0.2)	0.08 (0.04-0.2)	(0.1)	
Maze	1.7 (0.6-2.3)	1.2 (0.5-1.5)	0.8 (0.3-1.7)	(0.46)	
Ring remove	0.1 (0.03-0.3)	0 (0-0.05)	0.08 (0.03-0.2)	(0.15)	
Ring return	0.1 (0.07-0.3)	0.05 (0.02-0.1)	0.09 (0.02-0.12)	(0.45)	
Arc Cutting	0.4 (0.09-0.8)	0.6 (0.4-1.2)	0.3 (0.2-0.3)	(0.15)	
Diagnostic	0 (0-0)	0 (0-0.04)	0.01 (0-0.03)	(0.079)	
Meniscectomy	0 (0-0)	0.01 (0-0.03)	0.01 (0-0.03)	(0.16)	

Table C.8: Performance by group, measured by differences in hand dominance (% procedure; median and IQR) on each of the simulated tasks. Where p-values are in parentheses they do not reach corrected significance (<0.0071).

Task	Difference in hand dominance;% procedure, median (IQR)				p-value
	CCT	Intermediate	Novice	All-groups	
Probe	7.0 (3.6-24.5)	13.5 (9.5-20.2)	19.1 (9.0-26.0)	(0.5)	
Maze	27.1 (17.7-30.6)	31.2 (21.7-42.7)	33.9 (22.5-52.1)	(0.46)	
Ring remove	68.1 (59.1-73.4)	62.9 (53.5-75.9)	72.7 (70.6-78)	(0.23)	
Ring return	58.0 (44.3-76.6)	65.8 (55.6-69.6)	47.0 (25.4-62.9)	(0.15)	
Arc Cutting	67.8 (61.6-72.3)	64.0 (39.4-68.8)	46.7 (20.5-57.2)	(0.11)	
Diagnostic	18 (10.9-29.9)	32.0 (16.9-41.3)	16.6 (9.6-22.5)	(0.13)	
Menisectomy	60.4 (53.3-63.3)	44.1 (33.7-60.2)	25.4 (10.9-34.7)	(0.028)	

D

Randomised controlled trial to assess the
impact of objective motion performance
feedback on the learning curve in
simulated arthroscopy

Table D.1: Probability of an alarm for the LC-CUSUM at varying h thresholds at acceptable performance (‘consultant level’, μ_0) and unacceptable performance (μ_1 , or $\mu_0+20\%$) over 1000 simulations.

h	Hand movements performance		Minor movements performance		Smoothness performance		Time performance	
	μ_0	$\mu_0+20\%$	μ_0	$\mu_0+20\%$	μ_0	$\mu_0+20\%$	μ_0	$\mu_0+20\%$
0.1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	1.00	0.99	1.00	0.99	1.00	1.00	1.00	0.99
0.3	1.00	0.98	1.00	0.99	1.00	1.00	1.00	0.98
0.4	1.00	0.96	1.00	0.98	1.00	1.00	1.00	0.97
0.5	1.00	0.92	1.00	0.97	1.00	1.00	1.00	0.95
0.6	1.00	0.89	1.00	0.96	1.00	0.99	1.00	0.91
0.7	1.00	0.84	1.00	0.92	1.00	0.99	1.00	0.88
0.8	1.00	0.78	1.00	0.87	1.00	0.98	1.00	0.84
0.9	1.00	0.72	1.00	0.82	1.00	0.97	1.00	0.78
1	1.00	0.66	1.00	0.77	1.00	0.95	1.00	0.72
1.1	1.00	0.58	1.00	0.70	1.00	0.93	1.00	0.66
1.2	1.00	0.52	1.00	0.65	1.00	0.91	1.00	0.61
1.3	1.00	0.45	1.00	0.58	1.00	0.88	1.00	0.53
1.4	1.00	0.39	1.00	0.52	0.99	0.86	1.00	0.46
1.5	0.99	0.34	1.00	0.46	0.99	0.83	0.99	0.40
1.6	0.99	0.28	0.99	0.41	0.99	0.80	0.99	0.34
1.7	0.99	0.25	0.99	0.36	0.98	0.77	0.98	0.29
1.8	0.98	0.20	0.99	0.30	0.97	0.74	0.98	0.26
1.9	0.97	0.17	0.98	0.26	0.96	0.71	0.97	0.22
2	0.97	0.14	0.97	0.23	0.95	0.68	0.96	0.19
2.1	0.96	0.12	0.97	0.19	0.94	0.65	0.95	0.17
2.2	0.95	0.10	0.96	0.16	0.94	0.62	0.95	0.14
2.3	0.94	0.09	0.95	0.14	0.92	0.57	0.94	0.13
2.4	0.93	0.08	0.94	0.13	0.91	0.53	0.92	0.11
2.5	0.92	0.06	0.93	0.11	0.90	0.48	0.91	0.10
2.6	0.90	0.05	0.92	0.09	0.89	0.44	0.90	0.09
2.7	0.89	0.04	0.91	0.08	0.87	0.41	0.88	0.07
2.8	0.88	0.04	0.90	0.06	0.86	0.38	0.86	0.06
2.9	0.87	0.03	0.89	0.05	0.84	0.35	0.85	0.06
3	0.85	0.03	0.87	0.04	0.83	0.32	0.84	0.05
3.1	0.83	0.02	0.86	0.04	0.81	0.30	0.82	0.04
3.2	0.81	0.02	0.85	0.03	0.80	0.28	0.81	0.03
3.3	0.81	0.02	0.83	0.03	0.79	0.26	0.80	0.03
3.4	0.79	0.01	0.81	0.02	0.77	0.24	0.79	0.03
3.5	0.78	0.01	0.78	0.02	0.75	0.23	0.78	0.02
3.6	0.76	0.01	0.77	0.02	0.73	0.21	0.76	0.02
3.7	0.74	0.01	0.76	0.01	0.71	0.19	0.74	0.02
3.8	0.72	0.01	0.74	0.01	0.71	0.18	0.73	0.02
3.9	0.70	0.01	0.72	0.01	0.69	0.16	0.70	0.02
4	0.66	0.01	0.69	0.01	0.67	0.15	0.68	0.01

E

Randomised controlled trial to objectively assess the transfer of surgical simulation training to intra-operative surgical skills performance

E.1 Approved in-theatre protocol

Simulation-based Arthroscopic Surgery Programme

The study has approval from the Central Universities Research and Ethics Committee (MSD-IDREC-C1-2014-152) and NHS Trust Management Approval and Indemnity (Research and Development Reference: 11572). This study is registered at clinicaltrials.gov. Participants in this transfer validity study will complete one diagnostic round of a knee joint in theatre under supervision to investigate whether additional simulation training in addition to their junior orthopaedic posts improves their intra-operative performance. The participants will be SHO trainees who are either randomised to additional top-up simulation training during their orthopaedic rotation, or no additional simulation. WASP motion sensor output and arthroscopic video feeds will be recorded. Motion analysis metric results will be analysed, and simulation trained vs non-simulation trained groups will be compared.

Conduct of assessments

Operating list

The assessments will take place on the Thursday afternoon Acute Knee Arthroscopy lists. These are supervised by Mr Bottomley, Mr Deakin, Mr Jackson, and Prof Price. At present, no other lists are being used.

Pre-operative

Maximum of two study participants will attend any given list. The researcher (PGR) will attend the WHO briefing in the morning, introduce themselves and explain role during the list. They will not disclose which arm of the study the participants are in. The supervising consultant will confirm that the cases on the list are suitable for a diagnostic knee arthroscopy by a junior trainee (in terms of anatomy, patient health, time available on list etc) In advance of the arthroscopic cases, the researcher will set up equipment (laptop) in a corner of theatre in discussion with scrub, anaesthetic and surgical staff. Participants will attend for the afternoon session and introduce themselves to the scrub team, and be recorded as a new team member during the WHO timeout. Additional consent of the patient is not required by the ethical approval as there is no change in their care. The study investigates the effect of additional training during normal working activities in the Trust employee would permitted and expected to undertake. If more than one participant, the second will wait in the coffee room for the second case.

Operative

The supervising consultant will position the patient and perform the EUA, and give the participant the opportunity to do the same. The supervising consultant and participant will both attach the WASP sensors and the researcher will connect to the sensors remotely, before scrubbing up in the usual manner. There will be a spare sensor pair if required. The supervising consultant will WHO time out, and prep the surgical site. The researcher will connect to the consultant sensors, but not record data until performing the diagnostic arthroscopy. The consultant completes the planned procedure, and will assess whether the case is suitable for a diagnostic arthroscopy at the end of the case. The camera and scope should be

assembled, water running and stoppered off. The consultant inserts the cannula via the lateral portal into the supra-patellar pouch. Before inserting the scope, WASP data collection begins, then completes a diagnostic round according to the pre-agreed sequence, describing each step for the benefit of the participant. Data collection stops at the end of the round. Before the participant's attempt, the trochar will be inserted through the cannula into the supra-patellar pouch. Before removing the trochar and inserting the arthroscope, pause to begin WASP data streaming. The participant removes the trochar, inserts the arthroscope, and completes the round according to pre-agreed sequence, observed by the consultant. The consultant will assist with leg positioning (on hip/figure-of-four, as these cannot be simulated in the lab). When the diagnostic round is complete, the participant will hand the scope back to the consultant to conclude the procedure.

Knee arthroscopy diagnostic round (as per Howells et al 2008 and AJ Price et al 2015)

The previously agreed, validated diagnostic arthroscopy protocols from Howells et al 2008 [Transferring simulated arthroscopic skills into the operating theatre: a randomised study¹⁴⁴] and from AJ Price et al 2015 [Evidence-based surgical training in orthopaedics: how many arthroscopies of the knee are needed to achieve consultant level performance?⁸⁴] which took place at this centre were combined to be used as the training protocol for the trainees, and what they will be asked to replicate in theatre. The sequence will be slightly altered, where the final element (probing the patella) will become the first element due to the starting position during normal knee arthroscopy that was not replicable in the simulator.

START WASP RECORDING

Insertion

1. Scope inserted through lateral portal cannula into supra-patellar pouch with knee in extension.

Setting up

2. Orientation period: camera position, focus and general 3D orientation

Inspection phase - structure visualised to be deliberate focus of view

3. Identify trochlea. 4. Identify patella. Visualise all facets. 5. Swing back down medial gutter to enter medial compartment while flexing knee: supervising clinician to assist with positioning on hip. 6. Identify medial femoral condyle. 7. Identify medial tibial plateau. 8. Identify medial meniscus (posterior horn). 9. Enter intercondylar notch and identify anterior cruciate ligament. 10. Enter lateral compartment following intercondylar notch: supervising clinician to assist with positioning in figure-of-four. 11. Identify lateral femoral condyle, tibial plateau and meniscus (posterior horn).

Probing phase

12. Come back over into medial compartment again following intercondylar notch (supervising consultant to assist with positioning). 13. Insert probe under direct vision. 14. Probe articular surface of medial femoral condyle. 15. Probe weight-bearing surface of medial tibial plateau. 16. Probe and lift medial meniscus (posterior horn). 17. Enter intercondylar notch Probe anterior cruciate ligament. 18. Enter lateral compartment following intercondylar notch: supervising clinician to assist with positioning in figure-of-four. 19. Probe articular surface of lateral femoral condyle. 20. Probe weight-bearing surface of lateral tibial plateau. 21. Probe and lift lateral meniscus (posterior horn). NB: Supervisor to act as assistant, helping with positioning of the leg

END WASP RECORDING

Appropriateness of case for participant There may be some situations where the supervising consultant feel the case is regrettably not appropriate for the participant to perform a diagnostic round. These include, but are not limited to:

- Anatomic factors - the patients knee is too anatomically distorted to give an impression of a typical knee scope, e.g. intra-articular pathology, body habitus, limited range of flexion of the knee
- Anaesthetic factors - the patient is too unwell to prolong the case for training and the case must be completed urgently or aborted.

- Time factors - the list would become unmanageable if the trainee performed a part of the operation.

Interruptions During the case there may be cause, in the supervising clinicians judgement, to interrupt the participant that does not require the procedure to be aborted. The participant remains still during these. These include, but are not limited to:

- Lack of progress to next step
- Presence on intra-articular features that need consultant input e.g. joint lavage
- Repositioning of dislodged scope In these cases, the supervising clinician takes control of the procedure, while the participant remains still, elbows at side, hands held in front of chest so as to minimise hand movements.

Cause to abort procedure There may be cases where the supervising consultant feels that the participant should be interrupted and the case not continue for patient safety. The reasons will be recorded. These include, but are not limited to:

- Participant shows behaviour or attitude that is unacceptable to the supervising consultant
- Lack of progress to the extent that the consultant feels the procedure is unlikely to be completed by the participant, or would lead to an unacceptable increase in tourniquet time
- Participant persisting in deviating from agreed protocol despite being informed
- Supervising clinician recognises unsuitable training conditions or environment.

E.2 Simulated arthroscopy list and in-theatre performance results

Table E.1: Hand movements (n) during simulated arthroscopy list at baseline and follow-up. Unadjusted p-values and Holm-Bonferroni sequentially corrected p'-values after Mann-Whitney U test ($\alpha = 0.05$).

	Baseline				Follow-up							
	Control		Intervention		Control				Intervention			
	Median	IQR	Median	IQR	p-value	p'-value	Median	IQR	Median	IQR	p-value	p'-value
Probe	536	415-712	555	476-681	0.98	1	535	365-580	258	228-264	0.00025	0.0041
Maze	138	95-158	101	98-111	0.21	1	118	84-133	83	76-92	0.034	0.068
Flat ring remove	135	99-166	131	96-154	0.69	1	105	91-132	86	70-98	0.043	0.068
Flat ring return	197	131-243	181	137-210	0.68	1	147	105-186	94	76-117	0.0050	0.025
Arc cutting	281	234-345	246	205-339	0.56	1	195	166-279	104	91-130	0.00040	0.0056
Knee diagnostic	311	286-424	308	250-426	0.4	1	249	197-345	131	108-153	0.000034	0.00075
Knee meniscectomy	687	535-1071	640	501-745	0.48	1	574	431-814	290	262-310	0.000013	0.00031

Table E.2: Minor hand movements (n) during simulated arthroscopy list at baseline and follow-up. Unadjusted p-values and Holm-Bonferroni sequentially corrected p'-values after Mann-Whitney U test ($\alpha = 0.05$).

	Baseline				Follow-up							
	Control		Intervention		Control				Intervention			
	Median	IQR	Median	IQR	p-value	p'-value	Median	IQR	Median	IQR	p-value	p'-value
Probe	229	197-268	219	169-317	0.89	1	171	144-193	52	33-98	0.00099	0.01
Maze	194	154-238	182	155-207	0.54	1	180	129-274	110	94-133	0.0025	0.017
Flat ring remove	104	88-127	74	59-124	0.41	1	81	53-88	37	21-65	0.0074	0.03
Flat ring return	175	124-182	133	88-222	0.45	1	115	110-144	74	50-104	0.0021	0.017
Arc cutting	227	200-374	313	239-403	0.45	1	176	147-247	97	74-113	0.00019	0.0036
Knee diagnostic	113	70-154	99	81-132	0.91	1	62	43-125	19	10-28	0.00016	0.0034
Knee meniscectomy	275	209-297	302	198-386	0.62	1	227	138-249	98	75-116	0.00023	0.0041

Table E.3: Smoothness (ms^{-3}) during simulated arthroscopy list at baseline and follow-up. Unadjusted p-values and Holm-Bonferroni sequentially corrected p'-values after Mann-Whitney U test ($\alpha = 0.05$).

	Baseline			Follow-up								
	Control		Intervention		p-value	p'-value	Intervention					
	Median	IQR	Median	IQR			Median	IQR				
Probe	16241	12008-20652	15717	11886-20188	0.86	1	12528	8460-17907	6415	5983-8196	0.00093	0.01
Maze	6213	4442-6850	5327	3715-5556	0.12	1	5017	3776-5359	2769	2600-3235	0.00048	0.0062
Flat ring remove	4385	2985-5655	4261	3511-4593	0.48	1	3094	2842-4765	2268	2115-2538	0.00060	0.0072
Flat ring return	5834	4500-7269	5358	4322-6314	0.62	1	4157	3693-5212	2541	2282-2952	0.0000077	0.00019
Arc cutting	11765	7536-14784	11115	8960-12820	0.65	1	7012	5733-8876	3785	2989-4817	0.00018	0.0036
Knee diagnostic	10128	6985-12767	10703	7869-11541	0.95	1	6615	5556-11553	4282	3163-5237	0.0011	0.01
Knee meniscectomy	25232	15431-33528	19063	16725-27942	0.65	1	15842	13007-29573	9257	8398-10334	0.000031	0.00071

Table E.4: Time taken (s) during simulated arthroscopy list at baseline and follow-up. Unadjusted p-values and Holm-Bonferroni sequentially corrected p'-values after Mann-Whitney U test ($\alpha = 0.05$).

	Baseline			Follow-up						
	Intervention			Control			Intervention			
	Median	IQR		Median	IQR	p-value	Median	IQR	p-value	
Maze	154	118-188	147	131-159	147	121-187	97	86-108	0.00023	0.0041
Flat ring remove	120	92-142	110	85-133	90	77-113	63	46-74	0.010	0.03
Flat ring return	168	142-206	149	122-213	125	97-157	85	70-107	0.0021	0.017
Arc cutting	250	205-320	295	208-358	178	143-279	93	77-120	0.00023	0.0041
Knee diagnostic	211	187-248	221	168-241	132	111-214	71	60-83	0.00000020	0.0000056
Knee meniscectomy	449	342-605	471	336-551	331	272-466	165	154-190	0.0000014	0.000038

Table E.5: In-theatre performance data from participants, their supervisors, and their performance ratios during the same case (median and IQR). Unadjusted p-values and Holm-Bonferroni sequentially corrected p'-values after Mann-Whitney U test ($\alpha = 0.05$).

	Control group	Intervention group	p-value	p'-value
<i>Participant performance</i>				
Number of hand movements (n)	893 (747–1242)	544 (465–593)	0.00019	0.00048
Number of minor hand movements (n)	435 (310–652)	176 (133–209)	0.00016	0.00048
Smoothness (ms ⁻³)	36,846 (29,84–53,949)	25,842 (20,867–27,468)	0.00032	0.00048
Time taken (s)	573(477–860)	320 (294–392)	0.00002	0.00008
<i>Supervisor performance during same case</i>				
Number of hand movements (n)	333.5 (199–405.5)	277.5 (198–450)	1	1
Number of minor hand movements (n)	52 (26–77)	46.5 (34–69)	0.86	1
Smoothness (ms ⁻³)	17,059 (10,729–25,744)	16,834 (12,472–24,605)	0.59	1
Time taken (s)	165 (102–209)	141 (107–227)	0.94	1
<i>Performance ratio (participant:supervisor)</i>				
Relative number of hand movements	3.3 (2.2–4.8)	1.9 (1.5–2.1)	0.0091	0.018
Relative number of minor hand movements	10.3 (5.1–11.9)	3.8 (2.8–4.7)	0.017	0.018
Relative smoothness	2.6 (1.6–3.0)	1.2 (1.1–1.7)	0.0037	0.015
Relative time taken	4.3 (2.8–5.4)	2.1 (1.8–2.8)	0.0037	0.015

References

- [1] M C Lillie. “Cranial surgery dates back to mesolithic”. In: *Nature* 391.6670 (1998), pp. 854–854.
- [2] Harold Ellis. *A History of Surgery*. Cambridge University Press, Jan. 2002.
- [3] Julia Gresky et al. “New cases of trepanations from the 5th to 3rd millennia BC in Southern Russia in the context of previous research: Possible evidence for a ritually motivated tradition of cranial surgery?” In: *American Journal of Physical Anthropology* 160.4 (Apr. 2016), pp. 665–682.
- [4] Ján Roller. “Sushruta Samhita (Sushruta’s collection). Pioneers of plastic surgery”. In: *Plastic and Reconstructive Surgery* 76.4 (Oct. 1985), p. 667.
- [5] Katherine A Walker. “Pain and Surgery in England, circa 1620–circa 1740”. In: *Medical History* 59.02 (Mar. 2015), pp. 255–274.
- [6] Ritu Lakhtakia. “A Brief History of Breast Cancer: Part I: Surgical domination reinvented.” In: *Sultan Qaboos University Medical Journal* 14.2 (May 2014), e166–9.
- [7] I Loudon. “Why are (male) surgeons still addressed as Mr?” In: *BMJ* 321.7276 (Dec. 2000), pp. 1589–1591.
- [8] Humphry Davy. *Researches, Chemical and Philosophical; Chiefly Concerning Nitrous Oxide*. Or Dephlogisticated Nitrous Air and its Respiration. London: J Johnson, St. Paul’s Church-yard, 1800.
- [9] P Subramanian et al. “Orthopaedic surgeons: as strong as an ox and almost twice as clever? Multicentre prospective comparative study.” In: *BMJ* 343.dec15 1 (Dec. 2011), pp. d7506–d7506.
- [10] Atul Gawande. “Two hundred years of surgery.” In: *The New England Journal of Medicine* 366.18 (May 2012), pp. 1716–1723.
- [11] A Darzi, S Smith, and N Taffinder. “Assessing operative skill. Needs to become more objective.” In: *BMJ* 318.7188 (Apr. 1999), pp. 887–888.
- [12] Scott E Regenbogen et al. “Patterns of technical error among surgical malpractice claims: an analysis of strategies to prevent injury to surgical patients.” In: *Annals of Surgery* 246.5 (Nov. 2007), pp. 705–711.
- [13] L L Leape et al. “The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II.” In: *The New England Journal of Medicine* 324.6 (Feb. 1991), pp. 377–384.
- [14] Caprice C Greenberg et al. “Patterns of Communication Breakdowns Resulting in Injury to Surgical Patients”. In: *Journal of the American College of Surgeons* 204.4 (Apr. 2007), pp. 533–540.

- [15] Atul A Gawande et al. "Analysis of errors reported by surgeons at three teaching hospitals." In: *Surgery* 133.6 (June 2003), pp. 614–621.
- [16] Royal Australasian College of Surgeons. *Competencies*. URL: <https://www.surgeons.org/becoming-a-surgeon/surgical-education-training/competencies/>.
- [17] John F Seeley. *CanMEDS Framework*. URL: <http://www.royalcollege.ca/rcsite/canmeds/about-canmeds-e>.
- [18] S Swing. "ACGME launches outcomes assessment project". In: *Journal of the American Medical Association* 279.18 (1998), pp. 1492–1492.
- [19] Intercollegiate Surgical Curriculum Programme. *Intercollegiate Surgical Programme Trauma & Orthopaedic Surgery Syllabus 2017*. URL: https://www.iscp.ac.uk/curriculum/surgical/specialty_year_syllabus.aspx?enc=Alwrb1itzzAhusNAJUHPg==.
- [20] Ruth F McKee. "The Intercollegiate Surgical Curriculum Programme (ISCP)". In: *Surgery (Oxford)* 26.10 (Oct. 2008), pp. 411–416.
- [21] J Crossley et al. "Prospective observational study to evaluate NOTSS (Non-Technical Skills for Surgeons) for assessing trainees' non-technical performance in the operating theatre." In: *British Journal of Surgery* 98.7 (July 2011), pp. 1010–1020.
- [22] M T Nousiainen et al. "Resident education in orthopaedic trauma: the future role of competency-based medical education." In: *The Bone & Joint Journal* 98-B.10 (Oct. 2016), pp. 1320–1325.
- [23] Olympia Papachristofi. "Assessment of learning curves in complex surgical interventions: a consecutive case-series study". In: *Trials* (May 2016), pp. 1–10.
- [24] Nuffield Institute for Health, University of Leeds. "Hospital volume and health care outcomes, costs and patient access". In: *Effective Health Care* 2.8 (Dec. 1996), pp. 1–16.
- [25] John D Birkmeyer et al. "Surgical skill and complication rates after bariatric surgery." In: *The New England Journal of Medicine* 369.15 (Oct. 2013), pp. 1434–1442.
- [26] M J Reidy et al. "Do trainee surgeons have an adverse effect on the outcome after total hip arthroplasty? A TEN-YEAR REVIEW". In: *Journal of Bone and Joint Surgery, American Volume* 98B.3 (Mar. 2016), pp. 301–306.
- [27] Andrew J Pugely et al. "The Effect of Resident Participation on Short-term Outcomes After Orthopaedic Surgery". In: *Clinical Orthopaedics and Related Research®* 472.7 (Mar. 2014), pp. 2290–2300.
- [28] M Bridges and D L Diamond. "The financial impact of teaching surgical residents in the operating room". In: *The American Journal of Surgery* 177.1 (Jan. 1999), pp. 28–32.
- [29] L R Farnworth et al. "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." In: *The Iowa Orthopaedic Journal* 21 (2001), pp. 31–35.

- [30] P J Devereaux et al. "Need for expertise based randomised controlled trials". In: *BMJ* 330.7482 (2005), pp. 88–91.
- [31] Jonathan A Cook et al. "A systematic review of the use of an expertise-based randomised controlled trial design". In: *Trials* 16.1 (June 2015), pp. 1–10.
- [32] Peter McCulloch et al. "Surgical Innovation and Evaluation 3 No surgical innovation without evaluation: the IDEAL recommendations". In: *The Lancet* 374.9695 (Sept. 2009), pp. 1105–1112.
- [33] L P Fielding, S Stewart-Brown, and H A Dudley. "Surgeon-related variables and the clinical trial." In: *The Lancet* 2.8093 (Oct. 1978), pp. 778–779.
- [34] Jeffrey S Barkun et al. "Surgical Innovation and Evaluation 1 Evaluation and stages of surgical innovations". In: *The Lancet* 374.9695 (Sept. 2009), pp. 1089–1096.
- [35] T K Pandian et al. "At Home Preresidency Preparation for General Surgery Internship: A Pilot Study". In: *Journal of Surgical Education* (June 2017), pp. 1–6.
- [36] P Gogalniceanu, E Fitzgerald O'Connor, and A Raftery. "Undergraduate anatomy teaching in the UK". In: *The Bulletin of the Royal College of Surgeons of England* 91.3 (Mar. 2009), pp. 102–106.
- [37] Health Education England. *Broadening the Foundation Programme*. Tech. rep. Health Education England, Feb. 2014.
- [38] A Baskaradas, N Sivarasan, and M C Solan. "Finished Core Training but not 'Registrar Ready'? Making a Case for a Third Year of Core Surgical Training". In: *The Bulletin of the Royal College of Surgeons of England* 95.8 (Sept. 2013), pp. 1–5.
- [39] Matthew G Mullen et al. "Declining Operative Experience for Junior-Level Residents: Is This an Unintended Consequence of Minimally Invasive Surgery?" In: *Journal of Surgical Education* 73.4 (July 2016), pp. 609–615.
- [40] Brian A Weatherby et al. "The effect of resident work hour regulations on orthopaedic surgical education." In: *Journal of Surgical Orthopaedic Advances* 16.1 (2007), pp. 19–22.
- [41] Andrew Jenkinson. "Death of a proverb." In: *BMJ* 351 (Dec. 2015), h6226.
- [42] Maria Bussey and Griffiths Gareth. "Mentor, Examiner, Administrator: What kind of assessor are you?" In: *The Bulletin of the Royal College of Surgeons of England* 99.5 (May 2017), pp. 180–182.
- [43] GMC. "Training environments 2017: Key findings from the national training surveys". In: (Nov. 2017), pp. 1–52.
- [44] Anne-Lise D D'Angelo et al. "Residents' perception of skill decay during dedicated research time." In: *The Journal of Surgical Research* 199.1 (Nov. 2015), pp. 23–31.
- [45] Brian C George et al. "Readiness of US General Surgery Residents for Independent Practice." In: *Annals of Surgery* 266.4 (Oct. 2017), pp. 582–594.
- [46] Samer G Mattar et al. "General surgery residency inadequately prepares trainees for fellowship: results of a survey of fellowship program directors." In: *Annals of Surgery* 258.3 (Sept. 2013), pp. 440–449.

- [47] Emily M Bucholz et al. "Our trainees' confidence: results from a national survey of 4136 US general surgery residents." In: *Archives of Surgery* 146.8 (Aug. 2011), pp. 907–914.
- [48] Elspeth Hill et al. "You've got to know the rules to play the game: how medical students negotiate the hidden curriculum of surgical careers". In: *Medical Education* 48.9 (Aug. 2014), pp. 884–894.
- [49] P Manning. *Selection into ST3 training for Trauma and Orthopaedics for 2013*. URL: <http://www.boa.ac.uk/Publications/Documents/Proposal%20for%20selection%20into%20ST3.pdf>.
- [50] Anthony G Gallagher et al. "Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training." In: *Annals of Surgery* 241.2 (Feb. 2005), pp. 364–372.
- [51] James B Cowan et al. "Computer-Simulated Arthroscopic Knee Surgery: Effects of Distraction on Resident Performance". In: *Orthopedics* (Jan. 2016), pp. 1–6.
- [52] Benjamin M Artz, Amanda H Goodall, and Andrew J Oswald. "Boss Competence and Worker Well-Being". In: *ILR Review* 70.2 (July 2016), pp. 419–450.
- [53] Shakir Syed, Aun H Mirza, and Ashgar Ali. "A brief comparison of orthopaedic training in English-speaking countries." In: *Annals of the Royal College of Surgeons of England* 91.3 (Apr. 2009), pp. 226–231.
- [54] Andrew J Carr et al. "Advances in arthroscopy—indications and therapeutic applications". In: *Nature Reviews Rheumatology* 11.2 (Oct. 2014), pp. 77–85.
- [55] C W Kieser and R W Jackson. "Severin Nordentoft: The first arthroscopist." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 17.5 (May 2001), pp. 532–535.
- [56] R W Jackson. "Memories of the early days of arthroscopy: 1965-1975. The formative years." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 3.1 (1987), pp. 1–3.
- [57] David F Hamilton and Colin R Howie. "Knee arthroscopy: influence of systems for delivering healthcare on procedure rates." In: *BMJ* 351 (Sept. 2015), h4720.
- [58] A Judge et al. "Temporal trends and geographical variation in the use of subacromial decompression and rotator cuff repair of the shoulder in England." In: *The Bone & Joint Journal* 96-B.1 (Jan. 2014), pp. 70–74.
- [59] M Watanabe. "Memories of the early days of arthroscopy." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 2.4 (1986), pp. 209–214.
- [60] S S Jameson et al. "The burden of arthroscopy of the knee: a contemporary analysis of data from the English NHS." In: *The Journal of Bone and Joint Surgery. British volume* 93.10 (Oct. 2011), pp. 1327–1333.
- [61] Karen A Cullen, Margaret J Hall, and Aleksandr Golosinskiy. "Ambulatory surgery in the United States, 2006." In: *National Health Statistics Reports* 11 (Jan. 2009), pp. 1–25.
- [62] A G Gallagher et al. "An ergonomic analysis of the fulcrum effect in the acquisition of endoscopic skills." In: *Endoscopy* 30.7 (Sept. 1998), pp. 617–620.

- [63] Marie-Eve Lebel et al. "Observational Learning During Simulation-Based Training in Arthroscopy: Is It Useful to Novices?" In: *Journal of Surgical Education* (June 2017), pp. 1–9.
- [64] G J M Tuijthof et al. "Multirater agreement on arthroscopic image quality". In: *Proceedings of the Institution of Mechanical Engineers. Part H, Journal of engineering in medicine* 223.2 (Dec. 2008), pp. 179–187.
- [65] B Tang, G B Hanna, and A Cuschieri. "Analysis of errors enacted by surgical trainees during skills training courses." In: *Surgery* 138.1 (July 2005), pp. 14–20.
- [66] Aashay L Kekatpure et al. "Comparative analysis of visual field and image distortion in 30° and 70° arthroscopes". In: *Knee Surgery, Sports Traumatology, Arthroscopy* 24.7 (Nov. 2014), pp. 2359–2364.
- [67] Y Hui et al. "What skills should simulation training in arthroscopy teach residents? A focus on resident input." In: *International Journal of Computer Assisted Radiology and Surgery* 8.6 (Nov. 2013), pp. 945–953.
- [68] Rui M Duarte et al. "Benefits of radial distortion correction in arthroscopic surgery: a first experimental study on a knee model." In: *The International Journal of Medical Robotics and Computer Assisted Surgery* (Sept. 2014).
- [69] Adam J Spiers et al. "Negating the fulcrum effect in manual laparoscopic surgery: Investigating skill acquisition with a haptic simulator." In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 173.11 (May 2017), e1837.
- [70] G J M Tuijthof et al. "Performance of arthroscopic irrigation systems assessed with automatic blood detection". In: *Knee Surgery, Sports Traumatology, Arthroscopy* 19.11 (Apr. 2011), pp. 1948–1954.
- [71] George Chami et al. "Haptic Feedback Can Provide an Objective Assessment of Arthroscopic Skills". In: *Clinical Orthopaedics and Related Research* 466.4 (Jan. 2008), pp. 963–968.
- [72] Timothy B Frischmann, Mustapha Mouloua, and Katelyn Procci. "3-D gaming environment preferences: Inversion of the Y-axis". In: *Ergonomics* 58.11 (Sept. 2015), pp. 1792–1799.
- [73] Florian M Franzeck et al. "Prospective randomized controlled trial of simulator-based versus traditional in-surgery laparoscopic camera navigation training." In: *Surgical Endoscopy* 26.1 (Jan. 2012), pp. 235–241.
- [74] I R Crothers et al. "Experienced laparoscopic surgeons are automated to the "fulcrum effect": an ergonomic demonstration." In: *Endoscopy* 31.5 (June 1999), pp. 365–369.
- [75] Richard K Reznick and Helen MacRae. "Teaching surgical skills—changes in the wind." In: *The New England Journal of Medicine* 355.25 (Dec. 2006), pp. 2664–2669.
- [76] Carlos A Pellegrini. "Surgical Education in the United States". In: *Transactions of the Meeting of the American Surgical Association* 124 (2006), pp. 1–8.
- [77] P Singh and A Darzi. "Surgical training". In: *British Journal of Surgery* 100.3 (Jan. 2013), pp. 307–309.

- [78] Brian David Hodges. “A tea-steeping or i-Doc model for medical education?” In: *Academic Medicine* 85.9 Suppl (Sept. 2010), S34–44.
- [79] J Chikwe, A C de Souza, and J R Pepper. “No time to train the surgeons: More and more reforms result in less and less time for training”. In: *BMJ* 328.7437 (2004), pp. 418–419.
- [80] Joseph Bernstein. “Not the Last Word: Two Patients, Two Operating Rooms, One Surgeon—Does The Math Add Up?” In: *Clinical Orthopaedics and Related Research®* (July 2016), pp. 1–6.
- [81] Brian C George, Gary L Dunnington, and Debra A DaRosa. “Trainee Autonomy and Patient Safety”. In: *Annals of Surgery* (Nov. 2017), pp. 1–3.
- [82] Ben Chatterton. *The Spending Review: what does it mean for health and social care?* Tech. rep. London: The Health Foundation, Dec. 2015.
- [83] J G Penn-Barwell et al. “Is there variation in orthopaedic operative training experience between deaneries?” In: *Bone and Joint Journal Orthopaedic Proceedings Supplement* 97-B.SUPP 8 (June 2015), p. 22.
- [84] A J Price et al. “Evidence-based surgical training in orthopaedics: how many arthroscopies of the knee are needed to achieve consultant level performance?” In: *The Bone & Joint Journal* 97-B.10 (Oct. 2015), pp. 1309–1315.
- [85] Alice Miller and Julian Archer. “Impact of workplace based assessment on doctors’ education and performance: a systematic review.” In: *BMJ* 341 (Sept. 2010), p. c5064.
- [86] A Mathew, J D Beard, and M Bussey. “An evaluation of the use of direct observation of procedural skills in the intercollegiate surgical curriculum programme”. In: *Bulletin of The Royal College of Surgeons of England* 96.9 (Oct. 2014), pp. 10–13.
- [87] Krishna Moorthy et al. “Objective assessment of technical skills in surgery”. In: *BMJ* 327.7422 (Oct. 2003), pp. 1032–1037.
- [88] David A Rogers, Arthur S Elstein, and Georges Bordage. “Improving Continuing Medical Education for Surgical Techniques: Applying the Lessons Learned in the First Decade of Minimal Access Surgery”. In: *Annals of Surgery* 233.2 (Feb. 2001), pp. 159–166.
- [89] Simon Frostick. *Specialist Training in Trauma and Orthopaedics Curriculum August 2014*. Oct. 2013. URL: https://www.iscp.ac.uk/documents/syllabus_T0_2014.pdf.
- [90] Dimitrios Stefanidis. “Optimal Acquisition and Assessment of Proficiency on Simulators in Surgery”. In: *Surgical Clinics of North America* 90.3 (June 2010), pp. 475–489.
- [91] Mads Emil Jacobsen et al. “Testing Basic Competency in Knee Arthroscopy Using a Virtual Reality Simulator: Exploring Validity and Reliability.” In: *The Journal of Bone and Joint Surgery. American volume* 97.9 (May 2015), pp. 775–781.
- [92] K Anders Ericsson. “Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms”. In: *Intelligence* 45 (July 2014), pp. 81–103.

- [93] K Anders Ericsson, Ralf T Krampe, and Clemens Tesch-Römer. "The role of deliberate practice in the acquisition of expert performance." In: *Psychological Review* 100.3 (July 1993), pp. 363–406.
- [94] Roger L Kneebone. "Practice, rehearsal, and performance: an approach for simulation-based surgical and procedure training." In: *Journal of the American Medical Association* 302.12 (Sept. 2009), pp. 1336–1338.
- [95] L Cherkasskiy et al. "Patient-specific 3D models aid planning for triplane proximal femoral osteotomy in slipped capital femoral epiphysis." In: *Journal of Children's Orthopaedics* 11.2 (Apr. 2017), pp. 147–153.
- [96] Won Hyung A Ryu et al. "Systematic Review of Patient-Specific Surgical Simulation: Toward Advancing Medical Education." In: *Journal of Surgical Education* (June 2017).
- [97] Gamal Abdalla et al. "The effect of warm-up on surgical performance: a systematic review." In: *Surgical Endoscopy* 29.6 (June 2015), pp. 1259–1269.
- [98] Mustafa Karahan et al. *Effective Training of Arthroscopic Skills*. Ed. by Mustafa Karahan et al. 1st ed. Luxembourg: Springer Berlin Heidelberg, Apr. 2015.
- [99] D Brown. "The Role of Simulation in the Learning of Surgical Skills". In: *Bulletin of The Royal College of Surgeons of England* 95 (2013), pp. 292–295.
- [100] K Anders Ericsson. "Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains." In: *Academic Medicine* 79.10 Suppl (Oct. 2004), S70–81.
- [101] S Mackay et al. "Practice distribution in procedural skills training: a randomized controlled trial." In: *Surgical Endoscopy* 16.6 (June 2002), pp. 957–961.
- [102] Carol-Anne E Moulton et al. "Teaching surgical skills: what kind of practice makes perfect?: a randomized, controlled trial." In: *Annals of Surgery* 244.3 (Sept. 2006), pp. 400–409.
- [103] Anthony G Gallagher. "Metric-based simulation training to proficiency in medical education:- what it is and how to do it." In: *The Ulster Medical Journal* 81.3 (Sept. 2012), pp. 107–113.
- [104] G Wulf, C H Shea, and S Matschiner. "Frequent feedback enhances complex motor skill learning." In: *Journal of Motor Behavior* 30.2 (June 1998), pp. 180–192.
- [105] Jeremy Stoller et al. "Are There Detrimental Effects From Proficiency-Based Training in Fundamentals of Laparoscopic Surgery Among Novices? An Exploration of Goal Theory". In: *Journal of Surgical Education* 73.2 (Mar. 2016), pp. 215–221.
- [106] James R Korndorffer et al. "Simulator training for laparoscopic suturing using performance goals translates to the operating room." In: *Journal of the American College of Surgeons* 201.1 (July 2005), pp. 23–29.
- [107] Gabriele Wulf, Charles Shea, and Rebecca Lewthwaite. "Motor skill learning and performance: a review of influential factors." In: *Medical Education* 44.1 (Jan. 2010), pp. 75–84.
- [108] H R Champion and A G Gallagher. "Surgical simulation—a 'good idea whose time has come'". In: *British Journal of Surgery* 90.7 (July 2003), pp. 767–768.

- [109] Abtin Alvand et al. "Innate arthroscopic skills in medical students and variation in learning curves." In: *The Journal of Bone and Joint Surgery. American volume* 93.19 (Oct. 2011), e115(1–9).
- [110] A Alvand et al. "Identifying orthopaedic surgeons of the future: the inability of some medical students to achieve competence in basic arthroscopic tasks despite training: a randomised study." In: *The Journal of Bone and Joint Surgery. British volume* 93.12 (Dec. 2011), pp. 1586–1591.
- [111] C R Ramsay et al. "Statistical assessment of the learning curves of health technologies." In: *Health Technology Assessment* 5.12 (2001), pp. 1–79.
- [112] The Southern Surgeons Club. "A prospective analysis of 1518 laparoscopic cholecystectomies. The Southern Surgeons Club." In: *The New England Journal of Medicine* 324.16 (Apr. 1991), pp. 1073–1078.
- [113] Anthony G Gallagher et al. "Prospective, Randomized Assessment of Transfer of Training (ToT) and Transfer Effectiveness Ratio (TER) of Virtual Reality Simulation Training for Laparoscopic Skill Acquisition". In: *Annals of Surgery* 257.6 (June 2013), pp. 1025–1031.
- [114] Mahiben Maruthappu et al. "The Influence of Volume and Experience on Individual Surgical Performance". In: *Annals of Surgery* 261.4 (Apr. 2015), pp. 642–647.
- [115] Henry Marsh. "Better not look down..." In: *The Bulletin of the Royal College of Surgeons of England* 97.8 (Aug. 2015), pp. 339–342.
- [116] Teodor P Grantcharov and Richard K Reznick. "Teaching procedural skills." In: *BMJ* 336.7653 (May 2008), pp. 1129–1131.
- [117] F Nightingale. *Notes on nursing: what it is, and what it is not*. Boston: William Carter, 1860.
- [118] Matthew D Karam et al. "Current and future use of surgical skills training laboratories in orthopaedic resident education: a national survey." In: *The Journal of Bone and Joint Surgery. American volume* 95.1 (Jan. 2013), e4.
- [119] T W Pike et al. "Operating list composition and surgical performance." In: *British Journal of Surgery* 363 (Mar. 2018), p. 1.
- [120] Wenjing He et al. "The composition of surgical teams in the operating room and its impact on surgical team performance in China". In: *Surgical Endoscopy And Other Interventional Techniques* 28.5 (Dec. 2013), pp. 1473–1478.
- [121] M R de Leval et al. "Analysis of a cluster of surgical failures. Application to a series of neonatal arterial switch operations." In: *The Journal of Thoracic and Cardiovascular Surgery* 107.3 (Mar. 1994), 914–23– discussion 923–4.
- [122] Joint Committee on Surgical Training. *JCST Trauma & Orthopaedic SAC Newsletter Spring 2016*. May 2016. URL: <https://www.jcst.org/-/media/files/jcst/key-documents/newsletters/to-sac-newsletter-20161-spring.pdf>.
- [123] Joint Committee on Surgical Training. *JCST Trauma & Orthopaedic SAC Newsletter Summer 2014*. July 2014. URL: <https://www.jcst.org/-/media/files/jcst/archive/ortho-newsletters/to-sac-newsletter-summer-2014.pdf>.

- [124] P J O'Neill et al. "Arthroscopic proficiency: A survey of orthopaedic sports medicine fellowship directors and orthopaedic surgery department chairs". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 18.7 (Sept. 2002), pp. 795–800.
- [125] Michael Leonard et al. "Knee arthroscopy: how much training is necessary? A cross-sectional study ". In: *European Journal of Orthopaedic Surgery & Traumatology* 17.4 (Jan. 2007), pp. 359–362.
- [126] Dan Guttman et al. "Arthroscopic rotator cuff repair: the learning curve." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 21.4 (Apr. 2005), pp. 394–400.
- [127] Robert M Middleton et al. "Can Surgical Trainees Achieve Arthroscopic Competence at the End of Training Programs? A Cross-sectional Study Highlighting the Impact of Working Time Directives". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 33.6 (June 2017), pp. 1151–1158.
- [128] Richard L Angelo et al. "Metric Development for an Arthroscopic Bankart Procedure: Assessment of Face and Content Validity". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.8 (Aug. 2015), pp. 1430–1440.
- [129] Richard L Angelo et al. "A Proficiency-Based Progression Training Curriculum Coupled With a Model Simulator Results in the Acquisition of a Superior Arthroscopic Bankart Skill Set". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.10 (Oct. 2015), pp. 1854–1871.
- [130] Susan Jamieson. "Likert scales: how to (ab)use them." In: *Medical Education* 38.12 (Dec. 2004), pp. 1217–1218.
- [131] Rajesh Aggarwal et al. "An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room." In: *Annals of Surgery* 245.6 (June 2007), pp. 992–999.
- [132] Christopher M Brusalis et al. "Can a Novel, Low-Cost Simulation Model be Used to Teach Anterior Cruciate Ligament Graft Preparation?" In: *Journal of Pediatric Orthopaedics* 37.4 (June 2017), e277–e281.
- [133] Kashif Akhtar et al. "Training safer surgeons: How do patients view the role of simulation in orthopaedic training?" In: *Patient Safety in Surgery* 9.1 (2015), p. 11.
- [134] Paola Indovina et al. "Virtual Reality as a Distraction Intervention to Relieve Pain and Distress During Medical Procedures: A Comprehensive Literature Review." In: *The Clinical Journal of Pain* (Feb. 2018), p. 1.
- [135] G S Micheli. *Analysis of the Transfer of Training, Substitution, and Fidelity of Simulation of Transfer Equipment*. Tech. rep. 2. Orlando, FL 32813: United States Department of the Navy, June 1972.
- [136] Ryan Koehler et al. "Arthroscopic training resources in orthopedic resident education." In: *The Journal of Knee Surgery* 28.1 (Feb. 2015), pp. 67–74.
- [137] Mark A Vitale et al. "Training Resources in Arthroscopic Rotator Cuff Repair". In: *The Journal of Bone & Joint Surgery* 89.6 (June 2007).

- [138] Giovanna Ferrari. “PUBLIC ANATOMY LESSONS AND THE CARNIVAL: THE ANATOMY THEATRE OF BOLOGNA”. In: *Past & Present* 117.1 (Nov. 1987), pp. 50–106.
- [139] R Kyle Martin et al. “A Porcine Knee Model Is Valid for Use in the Evaluation of Arthroscopic Skills: A Pilot Study”. In: *Clinical Orthopaedics and Related Research®* 474.4 (Aug. 2015), pp. 1–6.
- [140] J Gilbody et al. “The use and effectiveness of cadaveric workshops in higher surgical training: a systematic review”. In: *Annals of the Royal College of Surgeons of England* 93.5 (July 2011), pp. 347–352.
- [141] R P Vance. “An introduction to the philosophical presuppositions of the animal liberation/rights movement.” In: *Journal of the American Medical Association* 268.13 (Oct. 1992), pp. 1715–1719.
- [142] Owen Dyer. “New York University medical cadavers found in mass grave”. In: *BMJ* 353 (2016), p. i3178.
- [143] David A Molho et al. “The Grapefruit: An Alternative Arthroscopic Tool Skill Platform.” In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 33.8 (Aug. 2017), pp. 1567–1572.
- [144] N R Howells et al. “Transferring simulated arthroscopic skills to the operating theatre: a randomised blinded study.” In: *The Journal of Bone and Joint Surgery. British volume* 90.4 (Apr. 2008), pp. 494–499.
- [145] W Dilworth Cannon et al. “Improving residency training in arthroscopic knee surgery with use of a virtual-reality simulator. A randomized blinded study.” In: *The Journal of Bone and Joint Surgery. American volume* 96.21 (Nov. 2014), pp. 1798–1806.
- [146] P Bradley and K Postlethwaite. “Setting up a clinical skills learning facility.” In: *Medical Education* 37 Suppl 1 (Nov. 2003), pp. 6–13.
- [147] Barbara Pitruzzella. *Commission on the Future of Surgery — Royal College of Surgeons*. 2017. URL: <https://www.rcseng.ac.uk/standards-and-research/future-of-surgery/>.
- [148] J Bruce Moseley et al. “A controlled trial of arthroscopic surgery for osteoarthritis of the knee.” In: *The New England Journal of Medicine* 347.2 (July 2002), pp. 81–88.
- [149] Raine Sihvonen et al. “Arthroscopic Partial Meniscectomy versus Sham Surgery for a Degenerative Meniscal Tear”. In: *The New England Journal of Medicine* 369.26 (Dec. 2013), pp. 2515–2524.
- [150] David J Beard et al. “Arthroscopic subacromial decompression for subacromial shoulder pain (CSAW): a multicentre, pragmatic, parallel group, placebo-controlled, three-group, randomised surgical trial.” In: *The Lancet* 391.10118 (Jan. 2018), pp. 329–338.
- [151] J E F Fitzgerald et al. “Improving the future of surgical training and education: Consensus recommendations from the Association of Surgeons in Training”. In: *International Journal of Surgery* 10.8 (2012), pp. 389–392.

- [152] J Temple. *Time for training*. A review of the impact of the European Working Time Directive on the quality of training. Health Education England: Health Education England, 2010.
- [153] Markku T Nousiainen et al. “Simulation for Teaching Orthopaedic Residents in a Competency-based Curriculum: Do the Benefits Justify the Increased Costs?” In: *Clinical Orthopaedics and Related Research*® (Sept. 2015), pp. 1–10.
- [154] Robert D Acton et al. “Unanticipated teaching demands rise with simulation training: strategies for managing faculty workload.” In: *Journal of Surgical Education* 72.3 (May 2015), pp. 522–529.
- [155] Michelle Meyer. “Two cheers for corporate experimentation: the A/B illusion and the virtues of data-driven innovation”. In: *Colorado Technology Law Journal* 13.2 (May 2015), pp. 273–332.
- [156] Viroj Wiwanitkit. “Computer-simulation training for arthroscopic surgery: what is to be considered?” In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 26.9 (Sept. 2010), 1146–author reply 1146–7.
- [157] Cynthia H Argani et al. “The case for simulation as part of a comprehensive patient safety program.” In: *American Journal of Obstetrics and Gynecology* 206.6 (June 2012), pp. 451–455.
- [158] Nasir I Bhatti, Aadil Ahmed, and Sukgi S Choi. “Identifying quality indicators of surgical training: A national survey.” In: *The Laryngoscope* 125.12 (Dec. 2015), pp. 2685–2689.
- [159] David A Cook and Thomas J Beckman. “Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application”. In: *The American Journal of Medicine* 119.2 (Feb. 2006), 166.e7–166.e16.
- [160] Iman Ghaderi et al. “Technical skills assessment toolbox: a review using the unitary framework of validity.” In: *Annals of Surgery* 261.2 (Feb. 2015), pp. 251–262.
- [161] D L Kirkpatrick and J Kirkpatrick. *Evaluating training programs*. American Society for Training and Development, 1975.
- [162] Ann Sofia S Thomsen et al. “Update on Simulation-Based Surgical Training and Assessment in Ophthalmology: A Systematic Review.” In: *Ophthalmology* (Apr. 2015).
- [163] R M Harden et al. “BEME Guide No. 1: Best evidence medical education”. In: *Medical Teacher* 21.6 (Nov. 1999), pp. 553–562.
- [164] Richard L Angelo. “Editorial Commentary: A Task May Be Done Quickly, but Not Necessarily Well”. In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 32.7 (July 2016), pp. 1335–1336.
- [165] Royal College of Surgeons. *Skills and Qualities of a Surgeon*. URL: <https://www.rcseng.ac.uk/careers-in-surgery/careers-support/skills-and-qualities-of-a-surgeon/>.
- [166] American Board of Orthopaedic Surgery. *Rules and Procedures for Residency Education, Part I, and Part II Examinations*. Sept. 2017.

- [167] Patrick L Ergina et al. “Surgical Innovation and Evaluation 2 Challenges in evaluating surgical innovation”. In: *The Lancet* 374.9695 (Sept. 2009), pp. 1097–1104.
- [168] Richard L Angelo. “Magellan and Copernicus: Arthroscopy Association of North America Seeking Excellence in Education”. In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.8 (Aug. 2015), pp. 1428–1429.
- [169] F Kiernan and F Rahman. “Measuring surgical performance: A risky game?” In: *The Surgeon* 13.4 (Aug. 2015), pp. 213–217.
- [170] Rajesh Aggarwal and Ara Darzi. “Technical-Skills Training in the 21st Century”. In: *The New England Journal of Medicine* 355.25 (Dec. 2006), pp. 2695–2696.
- [171] Joint Committee on Intercollegiate Examinations. “Intercollegiate Specialty Examination in Trauma & Orthopaedic Surgery Syllabus Blueprint 2016”. In: (Oct. 2016), pp. 1–23.
- [172] Royal Australasian College of Surgeons. *About the Fellowship Examination*. URL: <https://www.surgeons.org/becoming-a-surgeon/surgical-education-training/examinations/fellowship-examination/>.
- [173] James H Lubowitz et al. “Learning the Language of Copernicus.” In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.8 (Aug. 2015), pp. 1423–1425.
- [174] Royal College of Surgeons. *Avoiding unconscious bias*. Tech. rep. London: Royal College of Surgeons, May 2016.
- [175] C Rizan et al. “Are general surgeons able to accurately self-assess their level of technical skills?” In: *Annals of the Royal College of Surgeons of England* 97.8 (Nov. 2015), pp. 549–555.
- [176] Justin Kruger and David Dunning. “Unskilled and unaware of it: How difficulties in recognizing one’s own incompetence lead to inflated self-assessments.” In: *Journal of Personality and Social Psychology* 77.6 (Dec. 1999), pp. 1121–1134.
- [177] Alessandro Liberati et al. “The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration.” In: *BMJ* 339 (July 2009), b2700.
- [178] The Cochrane Collaboration. *Covidence*. 2018. URL: <http://community.cochrane.org/tools/review-production-tools/covidence>.
- [179] Kotaro Yamakado. “Quantification of the Learning Curve for Arthroscopic Suprascapular Nerve Decompression: An Evaluation of 300 Cases”. In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.2 (Feb. 2015), pp. 191–196.
- [180] Marc N Thomsen and Robert D Lang. “An experimental comparison of 3-dimensional and 2-dimensional endoscopic systems in a model.” In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 20.4 (Apr. 2004), pp. 419–423.
- [181] Justin Mitchell et al. “Endoscopic Trochanteric Bursectomy and Iliotibial Band Release for Persistent Trochanteric Bursitis”. In: *Arthroscopy Techniques* 5.5 (Oct. 2016), e1185–e1189.
- [182] Darcy A Reed et al. “Association between funding and quality of published medical education research.” In: *Journal of the American Medical Association* 298.9 (Sept. 2007), pp. 1002–1009.

- [183] J P T Higgins et al. "The Cochrane Collaboration's tool for assessing risk of bias in randomised trials". In: *BMJ* 343.oct18 2 (Oct. 2011), pp. d5928–d5928.
- [184] Brent A Ponce et al. "Telementoring: use of augmented reality in orthopaedic education: AAOS exhibit selection." In: *The Journal of Bone and Joint Surgery. American volume* 96.10 (May 2014), e84–e84.
- [185] Brian R Waterman et al. "Simulation Training Improves Surgical Proficiency and Safety During Diagnostic Shoulder Arthroscopy Performed by Residents." In: *Orthopedics* 39.3 (May 2016), e479–85.
- [186] John C Dunn et al. "Arthroscopic Shoulder Surgical Simulation Training Curriculum: Transfer Reliability and Maintenance of Skill Over Time". In: *Journal of Surgical Education* 72.6 (Aug. 2015), pp. 1118–1123.
- [187] P Strom et al. "Training in tasks with different visual-spatial components does not improve virtual arthroscopy performance." In: *Surgical Endoscopy* 18.1 (Jan. 2004), pp. 115–120.
- [188] Oleg Safir et al. "The benefits of Fundamentals of Laparoscopic Surgery (FLS) training on simulated arthroscopy performance." In: *Studies in Health Technology and Informatics* 173 (2012), pp. 412–417.
- [189] Kash Akhtar et al. "The Transferability of Generic Minimally Invasive Surgical Skills: Is There Crossover of Core Skills Between Laparoscopy and Arthroscopy?" In: *Journal of Surgical Education* 73.2 (Mar. 2016), pp. 329–338.
- [190] Terese T Horlocker et al. "Anesthetic, patient, and surgical risk factors for neurologic complications after prolonged total tourniquet time during total knee arthroplasty." In: *Anesthesia & Analgesia* 102.3 (Mar. 2006), pp. 950–955.
- [191] Brian J Daley et al. "How slow is too slow? Correlation of operative time to complications: an analysis from the Tennessee Surgical Quality Collaborative." In: *Journal of the American College of Surgeons* 220.4 (Apr. 2015), pp. 550–558.
- [192] Kash Akhtar et al. "Validation of a Virtual Reality Knee Arthroscopy Simulator for Surgical Training". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 29.10 (Oct. 2013), e138–e139.
- [193] Kash Akhtar et al. "Assessing Skills Decay on a Knee Arthroscopy Simulator". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 29.10 (Oct. 2013), e139.
- [194] A Alvand et al. "Simple visual parameters for objective assessment of arthroscopic skill." In: *The Journal of Bone and Joint Surgery. American volume* 94.13 (July 2012), e97.
- [195] Abtin Alvand et al. "Validating a global rating scale to monitor individual resident learning curves during arthroscopic knee meniscal repair." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 29.5 (May 2013), pp. 906–912.
- [196] Oke A Anakwenze et al. "Arthroscopic repair of large rotator cuff tears using the double-row technique: an analysis of surgeon experience on efficiency and outcomes". In: *Journal of Shoulder and Elbow Surgery* 22.1 (Jan. 2013), pp. 26–31.
- [197] Christine Andersen, Trine N Winding, and Martin S Vesterby. "Development of simulated arthroscopic skills." In: *Acta Orthopaedica* 82.1 (Feb. 2011), pp. 90–95.

- [198] Richard L Angelo et al. “The Bankart Performance Metrics Combined With a Shoulder Model Simulator Create a Precise and Accurate Training Tool for Measuring Surgeon Skill.” In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* (June 2015).
- [199] Richard L Angelo et al. “The Bankart Performance Metrics Combined With a Cadaveric Shoulder Create a Precise and Accurate Assessment Tool for Measuring Surgeon Skill”. In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.9 (Sept. 2015), pp. 1655–1670.
- [200] Sofia Bayona et al. “Assessment study of insight ARTHRO VR (®) arthroscopy virtual training simulator: face, content, and construct validities.” In: *Journal of Robotic Surgery* 2.3 (Sept. 2008), pp. 151–158.
- [201] James P Bliss, Hope S Hanner-Bailey, and Mark W Scerbo. “Determining the efficacy of an immersive trainer for arthroscopy skills.” In: *Studies in Health Technology and Informatics* 111 (2005), pp. 54–56.
- [202] Jonathan P Braman et al. “Development and validation of a basic arthroscopy skills simulator.” In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.1 (Jan. 2015), pp. 104–112.
- [203] Aaron Butler et al. “Do the skills acquired by novice surgeons using anatomic dry models transfer effectively to the task of diagnostic knee arthroscopy performed on cadaveric specimens?” In: *Journal of Bone and Joint Surgery, American Volume* 95.3 (Feb. 2013), e15(1–8).
- [204] C L Camp et al. “Improving Resident Performance in Knee Arthroscopy: A Prospective Value Assessment of Simulators and Cadaveric Skills Laboratories”. In: *The Journal of Bone & Joint Surgery* 98.3 (Feb. 2016), pp. 220–225.
- [205] W Dilworth Cannon et al. “Evaluation of skill level between trainees and community orthopaedic surgeons using a virtual reality arthroscopic knee simulator.” In: *The Journal of Bone and Joint Surgery. American volume* 96.7 (Apr. 2014), e57.
- [206] R Castricini et al. “Arthroscopic Latarjet procedure: analysis of the learning curve”. In: *Musculoskeletal Surgery* 97.S1 (Apr. 2013), pp. 93–98.
- [207] Justues Chang et al. “Global Rating Scales and Motion Analysis Are Valid Proficiency Metrics in Virtual and Benchtop Knee Arthroscopy Simulators”. In: *Clinical Orthopaedics and Related Research®* (Aug. 2015), pp. 1–9.
- [208] G Cunningham et al. “Comparison of arthroscopic and open Latarjet with a learning curve analysis.” In: *Knee Surgery, Sports Traumatology, Arthroscopy* 24.2 (Feb. 2016), pp. 540–545.
- [209] Francisco del Pinal et al. “Early Experience With (Dry) Arthroscopic 4-Corner Arthrodesis: From a 4-Hour Operation to a Tourniquet Time”. In: *The Journal of Hand Surgery* 37.11 (Nov. 2012), pp. 2389–2399.
- [210] Michael J Elliott et al. “Diagnostic knee arthroscopy: a pilot study to evaluate surgical skills.” In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 28.2 (Feb. 2012), pp. 218–224.

- [211] Abelardo Escoto et al. "Force sensing-based simulator for arthroscopic skills assessment in orthopaedic knee surgery." In: *Studies in Health Technology and Informatics* 173 (2012), pp. 129–135.
- [212] Abelardo Escoto et al. "A knee arthroscopy simulator: design and validation." In: *Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society* 2013 (2013), pp. 5715–5718.
- [213] Sandro F Fucentese et al. "Evaluation of a virtual-reality-based simulator using passive haptic feedback for knee arthroscopy." In: *Knee Surgery, Sports Traumatology, Arthroscopy* 23.4 (Apr. 2015), pp. 1077–1085.
- [214] Maulik J Gandhi, Michael J Anderton, and Lennard Funk. "Arthroscopic Skills Acquisition Tools: An Online Simulator for Arthroscopy Training." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.9 (June 2015), pp. 1671–1679.
- [215] Maulik J Gandhi, Michael J Anderton, and Lennard Funk. "Subjective assessment of hand-eye coordination and manual dexterity does not accurately predict performance on a virtual reality shoulder arthroscopy simulator". In: *Société Européenne pour la Chirurgie de l'Epaule et du Coude 2015*. 2015.
- [216] Brian B Gilmer et al. "Orthopaedic Residents Improve Confidence and Knot-Tying Speed With a Skills Course". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.7 (July 2015), 1343–1348.e2.
- [217] Andreas H Gomoll et al. "Surgical experience correlates with performance on a virtual reality simulator for shoulder arthroscopy." In: *The American Journal of Sports Medicine* 35.6 (June 2007), pp. 883–888.
- [218] Andreas H Gomoll et al. "Individual skill progression on a virtual reality simulator for shoulder arthroscopy: a 3-year follow-up study." In: *The American Journal of Sports Medicine* 36.6 (June 2008), pp. 1139–1142.
- [219] R Frank Henn et al. "Shoulder arthroscopy simulator training improves shoulder arthroscopy performance in a cadaveric model." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 29.6 (June 2013), pp. 982–985.
- [220] Nick R Howells et al. "Motion analysis: a validated method for showing skill levels in arthroscopy." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 24.3 (Mar. 2008), pp. 335–342.
- [221] N R Howells et al. "Retention of arthroscopic shoulder skills learned with use of a simulator. Demonstration of a learning curve and loss of performance level after a time delay." In: *The Journal of Bone and Joint Surgery. American volume* 91.5 (May 2009), pp. 1207–1213.
- [222] Sung-Jae Kim et al. "Arthroscopic treatment for limitation of motion of the elbow: the learning curve". In: *Knee Surgery, Sports Traumatology, Arthroscopy* 19.6 (Dec. 2010), pp. 1013–1018.
- [223] Georgina S J Kirby et al. "Assessing Arthroscopic Skills Using Wireless Elbow-Worn Motion Sensors". In: *The Journal of Bone and Joint Surgery. American volume* 97.13 (July 2015), pp. 1119–1127.

- [224] Ryan J Koehler et al. "Assessing Diagnostic Arthroscopy Performance in the Operating Room Using the Arthroscopic Surgery Skill Evaluation Tool (ASSET)." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.12 (Dec. 2015), 2314–2319.e2.
- [225] Sujith Konan, Shin-Jae Rhee, and Fares S Haddad. "Hip arthroscopy: analysis of a single surgeon's learning experience." In: *Journal of Bone and Joint Surgery, American Volume* 93.Supplement 2 (May 2011), pp. 52–56.
- [226] Young-Kyun Lee et al. "Learning curve of acetabular cup positioning in total hip arthroplasty using a cumulative summation test for learning curve (LC-CUSUM)." In: *The Journal of Arthroplasty* 29.3 (Mar. 2014), pp. 586–589.
- [227] Kevin D Martin et al. "Arthroscopic basic task performance in shoulder simulator model correlates with similar task performance in cadavers." In: *The Journal of Bone and Joint Surgery. American volume* 93.21 (Nov. 2011), e1271–5.
- [228] Kevin D Martin et al. "Shoulder arthroscopy simulator performance correlates with resident and shoulder arthroscopy experience." In: *The Journal of Bone and Joint Surgery. American volume* 94.21 (Nov. 2012), e160.
- [229] Kevin D Martin et al. "Ankle Arthroscopy Simulation Improves Basic Skills, Anatomic Recognition, and Proficiency During Diagnostic Examination of Residents in Training." In: *Foot and Ankle International* 36.7 (July 2015), pp. 827–835.
- [230] Avril D McCarthy et al. "Passive haptics in a knee arthroscopy simulator: is it valid for core skills training?" In: *Clinical Orthopaedics and Related Research®* 442 (Jan. 2006), pp. 13–20.
- [231] R M Middleton et al. "Which Global Rating Scale?: A Comparison of the ASSET, BAKSSS, and IGARS for the Assessment of Simulated Arthroscopic Skills". In: *The Journal of Bone & Joint Surgery* 98.1 (Jan. 2016), pp. 75–81.
- [232] G T Nicandri et al. "Assessing Improvement in Diagnostic Knee Arthroscopic Skill During the AAOS Fundamentals of Knee and Shoulder Arthroscopy Course for Orthopaedic Residents". In: *Orthopaedic Journal of Sports Medicine* 3.2 Suppl (Aug. 2015), 2325967115S00061.
- [233] Robert A Pedowitz, James Esch, and Steve Snyder. "Evaluation of a virtual reality simulator for arthroscopy skills development." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 18.6 (July 2002), E29.
- [234] Thomas C B Pollard et al. "Simulated hip arthroscopy skills: learning curves with the lateral and supine patient positions: a randomized trial." In: *The Journal of Bone and Joint Surgery. American volume* 94.10 (May 2012), e68.
- [235] Stefan Rahm et al. "Performance of medical students on a virtual reality simulator for knee arthroscopy: an analysis of learning curves and predictors of performance". In: *BMC Surgery* 16.1 (Mar. 2016), pp. 1–8.
- [236] Brian J Rebolledo et al. "Arthroscopy skills development with a surgical simulator: a comparative study in orthopaedic surgery residents." In: *The American Journal of Sports Medicine* 43.6 (June 2015), pp. 1526–1529.

- [237] Kelsey Rose and Robert Pedowitz. “Fundamental arthroscopic skill differentiation with virtual reality simulation.” In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.2 (Feb. 2015), pp. 299–305.
- [238] S Smith et al. “Early experience and validation work with Procedicus VA—the Prosolvia virtual reality shoulder arthroscopy trainer.” In: *Studies in Health Technology and Informatics* 62 (1999), pp. 337–43.
- [239] Sakti Srivastava et al. “Initial evaluation of a shoulder arthroscopy simulator: establishing construct validity.” In: *Journal of shoulder and elbow surgery / American Shoulder and Elbow Surgeons ... [et al.]* 13.2 (Mar. 2004), pp. 196–205.
- [240] J J Stunt et al. “Validation of the ArthroS virtual reality simulator for arthroscopic skills.” In: *Knee Surgery, Sports Traumatology, Arthroscopy* 23.11 (Nov. 2015), pp. 3436–3442.
- [241] J J Stunt et al. “Validation of the PASSPORT V2 training environment for arthroscopic skills.” In: *Knee Surgery, Sports Traumatology, Arthroscopy* (Aug. 2014), pp. 1–8.
- [242] Yasutaka Tashiro et al. “Evaluation of skills in arthroscopic training based on trajectory and force data.” In: *Clinical Orthopaedics and Related Research®* 467.2 (Feb. 2009), pp. 546–552.
- [243] Gabriëlle J M Tuijthof et al. “First validation of the PASSPORT training environment for arthroscopic skills.” In: *Knee Surgery, Sports Traumatology, Arthroscopy* 18.2 (Feb. 2010), pp. 218–224.
- [244] Gabriëlle J M Tuijthof et al. “Does perception of usefulness of arthroscopic simulators differ with levels of experience?” In: *Clinical Orthopaedics and Related Research®* 469.6 (June 2011), pp. 1701–1708.
- [245] Bing Wang et al. “An evaluation of the learning curve for a complex surgical technique: the full endoscopic interlaminar approach for lumbar disc herniations.” In: *The Spine Journal* 11.2 (Feb. 2011), pp. 122–130.
- [246] Ivan H Wong et al. “Systematic instruction of arthroscopic knot tying with the ArK Trainer: an objective evaluation tool”. In: *Knee Surgery, Sports Traumatology, Arthroscopy* 23.3 (June 2013), pp. 912–918.
- [247] W F M Jackson et al. “Learning and retaining simulated arthroscopic meniscal repair skills.” In: *The Journal of Bone and Joint Surgery. American volume* 94.17 (Sept. 2012), e132.
- [248] Stuart E Dreyfus and Hubert L Dreyfus. *A Five-stage Model of the Mental Activities Involved in Directed Skill Acquisition*. Operations Research Center, USAF. UC Berkley, Feb. 1980.
- [249] R J Koehler et al. “The Arthroscopic Surgical Skill Evaluation Tool (ASSET)”. In: *The American Journal of Sports Medicine* 41.6 (June 2013), pp. 1229–1237.
- [250] Aaron Insel et al. “The Development of an Objective Model to Assess Arthroscopic Performance”. In: *The Journal of Bone & Joint Surgery* 91.9 (Sept. 2009), pp. 2287–10.
- [251] Antonia C Hoyle et al. “Validation of a global rating scale for shoulder arthroscopy: a pilot study”. In: *Shoulder & Elbow* 4.1 (Nov. 2011), pp. 16–21.

- [252] Jesse A Slade Shantz et al. "Validation of a global assessment of arthroscopic skills in a cadaveric knee model." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 29.1 (Jan. 2013), pp. 106–112.
- [253] GMC. *Survey design - best practice guidelines*. URL: http://www.gmc-uk.org/NTS_2015_Briefing_note_1_Annex_A__Best_practice_guidelines_.pdf_57452861.pdf.
- [254] R Reznick et al. "Testing technical skill via an innovative "bench station" examination." In: *The American Journal of Surgery* 173.3 (Mar. 1997), pp. 226–230.
- [255] Sofia Bayona et al. "Assessing Performance in Shoulder Arthroscopy: The Imperial Global Arthroscopy Rating Scale (IGARS)." In: *The Journal of Bone and Joint Surgery. American volume* 96.13 (July 2014), e112–e112.
- [256] T Dwyer et al. "Simulation of Anterior Cruciate Ligament Reconstruction in a Dry Model". In: *The American Journal of Sports Medicine* 43.12 (Nov. 2015), pp. 2997–3004.
- [257] Tim Dwyer et al. "Use of an Objective Structured Assessment of Technical Skill (OSATS) After a Sports Medicine Rotation". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* (July 2016).
- [258] Justin Hodgins et al. "The knee arthroscopy learning curve: quantitative assessment of surgical skills." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 30.5 (May 2014), pp. 613–621.
- [259] Ryan J Koehler and Gregg T Nicandri. "Using the arthroscopic surgery skill evaluation tool as a pass-fail examination." In: *The Journal of Bone and Joint Surgery. American volume* 95.23 (Dec. 2013), e1871–6.
- [260] Tyson Olson et al. "Is there a Valid and Reliable Assessment of Diagnostic Knee Arthroscopy Skill?" In: *Clinical Orthopaedics and Related Research®* 471.5 (Dec. 2012), pp. 1670–1676.
- [261] James H Lubowitz. "Editorial Commentary: Arthroscopic Training and Proficiency Assessment Are Hot Topics". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.8 (Aug. 2015), pp. 1450–1451.
- [262] G Regehr et al. "Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination." In: *Academic Medicine* 73.9 (Sept. 1998), pp. 993–997.
- [263] Christopher L Talbot et al. "The Shoulder Objective Practical Assessment Tool: Evaluation of a New Tool Assessing Residents Learning in Diagnostic Shoulder Arthroscopy". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.8 (Aug. 2015), pp. 1441–1449.
- [264] Robert Langner and Simon B Eickhoff. "Sustaining attention to simple tasks: A meta-analytic review of the neural mechanisms of vigilant attention." In: *Psychological Bulletin* 139.4 (2013), pp. 870–900.
- [265] Robert A Pedowitz et al. "Objective Assessment of Knot-Tying Proficiency With the Fundamentals of Arthroscopic Surgery Training Program Workstation and Knot Tester." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 31.10 (Oct. 2015), pp. 1872–1879.

- [266] Pemra C Unalan et al. "A basic arthroscopy course based on motor skill training." In: *Knee Surgery, Sports Traumatology, Arthroscopy* 18.10 (Oct. 2010), pp. 1395–1399.
- [267] B Hodges, G Regehr, and D Martin. "Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it." In: *Academic Medicine* 76.10 Suppl (Oct. 2001), S87–9.
- [268] Ryan P Coughlin et al. "A Validated Orthopaedic Surgical Simulation Model for Training and Evaluation of Basic Arthroscopic Skills." In: *The Journal of Bone and Joint Surgery. American volume* 97.17 (Sept. 2015), pp. 1465–1471.
- [269] Eric Ferkel et al. "Arthroscopic Shoulder Simulation: Can a Computer Perceive Expertise? (SS-61)". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 29.6 (June 2013), e29.
- [270] Kashif Akhtar et al. "Training safer orthopedic surgeons. Construct validation of a virtual-reality simulator for hip fracture surgery." In: *Acta Orthopaedica* 86.5 (2015), pp. 616–621.
- [271] Caroline E Gaymer et al. "Fluoroscopic Radiation Exposure During Hip Arthroscopy". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 29.5 (May 2013), pp. 870–873.
- [272] Farshad Adib et al. "Paper 14: Duration Of Fluoroscopic Exposure In Hip Arthroscopy". In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 28.6 (June 2012), e51.
- [273] Young-Kyun Lee et al. "Learning curve of basic hip arthroscopy technique: CUSUM analysis." In: *Knee Surgery, Sports Traumatology, Arthroscopy* 21.8 (Aug. 2013), pp. 1940–1944.
- [274] Justin Hodgins and Christian Veillette. "Arthroscopic proficiency: methods in evaluating competency". In: *Bmc Medical Education* 13.1 (May 2013), p. 61.
- [275] Rachel M Frank et al. "Utility of modern arthroscopic simulator training models." In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 30.1 (Jan. 2014), pp. 121–133.
- [276] V Datta et al. "Motion analysis in the assessment of surgical skill". In: *Computer Methods in Biomechanics and Biomedical Engineering* 4 (Feb. 2001), pp. 515–523.
- [277] Hiroyuki Egi et al. "Evaluating the correlation between the HUESAD and OSATS scores: Concurrent validity study". In: *Minimally Invasive Therapy* 22.3 (July 2013), pp. 144–149.
- [278] Sif H Arnold et al. "Three-dimensional motion tracking correlates with skill level in upper gastrointestinal endoscopy." In: *Endoscopy* 47.9 (Sept. 2015), pp. 825–828.
- [279] M K Chmarra et al. "The influence of experience and camera holding on laparoscopic instrument movements measured with the TrEndo tracking system." In: *Surgical Endoscopy* 21.11 (Nov. 2007), pp. 2069–2075.
- [280] Yaron Munz et al. "Ceiling effect in technical skills of surgical residents." In: *The American Journal of Surgery* 188.3 (Sept. 2004), pp. 294–300.

- [281] Patrick Garfield Roberts et al. “Validation of the updated ArthroS simulator: face and construct validity of a passive haptic virtual reality simulator with novel performance metrics”. In: *Knee Surgery, Sports Traumatology, Arthroscopy* 25.2 (Apr. 2016), pp. 616–625.
- [282] Red Shift Labs. *Using Accelerometers to Estimate Position and Velocity*. URL: <http://www.chrobotics.com/library/accel-position-velocity>.
- [283] Jamie Ferguson et al. “Newly acquired arthroscopic skills: Are they transferable during simulator training of other joints?” In: *Knee Surgery, Sports Traumatology, Arthroscopy* 25.2 (Aug. 2015), pp. 1–8.
- [284] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017.
- [285] J Martin Bland and Douglas G Altman. “Correlation in restricted ranges of data.” In: *BMJ* 342 (Mar. 2011), p. d556.
- [286] Rajesh Aggarwal et al. “Proving the effectiveness of virtual reality simulation for training in laparoscopic surgery.” In: *Annals of Surgery* 246.5 (Nov. 2007), pp. 771–779.
- [287] Polhemus. *Polhemus PATRIOT Wireless*. URL: <http://polhemus.com/motion-tracking/all-trackers/patriot-wireless>.
- [288] Axivity. *WAX9 Data Sheet*. Sept. 2015. URL: <https://axivity.com/downloads/wax9>.
- [289] Uli Mezger, Claudia Jendrewski, and Michael Bartels. “Navigation in surgery.” In: *Langenbeck’s Archives of Surgery* 398.4 (Apr. 2013), pp. 501–514.
- [290] Robert A Pedowitz. “Editorial Commentary: Improving Proficiency in Arthroscopic Knot Tying: Focus on Performance Over Appearance”. In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* 32.7 (July 2016), pp. 1400–1401.
- [291] Gunnar Ahlberg et al. “Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies.” In: *American Journal of Surgery* 193.6 (June 2007), pp. 797–804.
- [292] British Orthopaedic Association. *Trauma and Orthopaedic Curriculum Mapped to Simulation Options*. 2013. URL: https://www.iscp.ac.uk/static/public/T0_Putting_simulation_into_practice.pdf.
- [293] Robert A Pedowitz et al. *Sawbones FAST Arthroscopy Training System*. Feb. 2017.
- [294] J M Bland and D G Altman. “Multiple significance tests: the Bonferroni method.” In: *BMJ* 310.6973 (Jan. 1995), p. 170.
- [295] Sanghoon Chae, Sung-Weon Jung, and Hyung-Soon Park. “In vivo biomechanical measurement and haptic simulation of portal placement procedure in shoulder arthroscopic surgery.” In: *PLoS ONE* 13.3 (2018), e0193736.
- [296] Gregory Lopez et al. “A Cost-Effective Junior Resident Training and Assessment Simulator for Orthopaedic Surgical Skills via Fundamentals of Orthopaedic Surgery: AAOS Exhibit Selection.” In: *The Journal of Bone and Joint Surgery. American volume* 97.8 (Apr. 2015), pp. 659–666.

- [297] Takahisa Suzuki et al. “An evaluation of the endoscopic surgical skills assessment using a video analysis software program”. In: *Surgical Endoscopy* 29.7 (Oct. 2014), pp. 1804–1808.
- [298] Anne-Lise D D’Angelo et al. “Idle time: an underdeveloped performance metric for assessing surgical skill.” In: *American Journal of Surgery* 209.4 (Apr. 2015), pp. 645–651.
- [299] Nick R Parsons et al. “An evaluation of the quality of statistical design and analysis of published medical research: results from a systematic survey of general orthopaedic journals.” In: *BMC Medical Research Methodology* 12.1 (Apr. 2012), p. 60.
- [300] Arthur Gilbert Bills. *General Experimental Psychology*. Longmans, Green and Co, 1934.
- [301] Jonathan A Cook, Craig R Ramsay, and Peter Fayers. “Using the literature to quantify the learning curve: a case study.” In: *International Journal of Technology Assessment in Health Care* 23.2 (2007), pp. 255–260.
- [302] Accrediation Council for Graudate Medical Education. *Orthopaedic Surgery Minimum Numbers*. Sept. 2014. URL: https://www.acgme.org/Portals/0/PFAssets/ProgramResources/260_ORs_Case_Log_Minimum_Numbers.pdf.
- [303] Joint Committee on Surgical Training. *Certification guidelines for Trauma and Orthopaedic Surgery 2017*. Aug. 2017. URL: <https://www.jcst.org/quality-assurance/certification-guidelines-and-checklists/>.
- [304] R A Schmidt et al. “Summary knowledge of results for skill acquisition: support for the guidance hypothesis.” In: *Journal of Experimental Psychology. Learning, Memory, and Cognition* 15.2 (Mar. 1989), pp. 352–359.
- [305] Dimitrios Stefanidis, Daniel J Scott, and James R Korndorffer. “Do metrics matter? Time versus motion tracking for performance assessment of proficiency-based laparoscopic skills training.” In: *Simulation in Healthcare* 4.2 (2009), pp. 104–108.
- [306] Gideon Sroka et al. “Fundamentals of laparoscopic surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial.” In: *American Journal of Surgery* 199.1 (Jan. 2010), pp. 115–120.
- [307] Teodor P Grantcharov, Svend Schulze, and Viggo B Kristiansen. “The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room”. In: *Surgical Endoscopy And Other Interventional Techniques* 21.12 (Apr. 2007), pp. 2240–2243.
- [308] Gyusung I Lee and Mija R Lee. “Can a virtual reality surgical simulation training provide a self-driven and mentor-free skills learning? Investigation of the practical influence of the performance metrics from the virtual reality robotic surgery simulator on the skill learning and associated cognitive workloads.” In: *Surgical Endoscopy* 7.5 (June 2017), p. 431.
- [309] D J Biau et al. “Quantitative and individualized assessment of the learning curve using LC-CUSUM”. In: *British Journal of Surgery* 95.7 (July 2008), pp. 925–929.

- [310] D Biau and R Porcher. “LC-CUSUM test.” In: *Anaesthesia* 65.7 (July 2010), pp. 755–756.
- [311] D J Biau, R Porcher, and L J Salomon. “CUSUM: a tool for ongoing assessment of performance”. In: *Ultrasound in Obstetrics and Gynecology* 31.3 (Mar. 2008), pp. 252–255.
- [312] T O Lim et al. “Assessing doctors’ competence: application of CUSUM technique in monitoring doctors’ performance”. In: *International Journal for Quality in Health Care* 14.3 (June 2002), pp. 251–258.
- [313] O A Grigg, V T Farewell, and D J Spiegelhalter. “Use of risk-adjusted CUSUM and RSPRT charts for monitoring in medical contexts”. In: *Statistical Methods in Medical Research* 12.2 (Mar. 2003), pp. 147–170.
- [314] David J Biau and Raphaël Porcher. “A method for monitoring a process from an out of control to an in control state: Application to the learning curve.” In: *Statistics in Medicine* 29.18 (Aug. 2010), pp. 1900–1909.
- [315] Robert Pedowitz and David F Martin. *ABOS Curriculum Module 14: Basic Arthroscopy Skills*. Nov. 2013. URL: https://www.abos.org/media/7061/module_14_basic_arthroscopy_skills_edit.pdf.
- [316] Liane S Feldman et al. “A method to characterize the learning curve for performance of a fundamental laparoscopic simulator task: defining "learning plateau" and "learning rate".” In: *Surgery* 146.2 (Aug. 2009), pp. 381–386.
- [317] Mark R Hutchinson. *Fundamentals of Arthroscopic Surgery Training (FAST) Curriculum*. June 2013. URL: https://www.abos.org/media/7061/module_14_basic_arthroscopy_skills_edit.pdf.
- [318] S A Fraser et al. “Characterizing the learning curve for a basic laparoscopic drill.” In: *Surgical Endoscopy* 19.12 (Dec. 2005), pp. 1572–1578.
- [319] Maximilian J Johnston et al. “An overview of research priorities in surgical simulation: what the literature shows has been achieved during the 21st century and what remains.” In: *American Journal of Surgery* 211.1 (Jan. 2016), pp. 214–225.
- [320] S de Montbrun, L Satterthwaite, and T P Grantcharov. “Setting pass scores for assessment of technical performance by surgical trainees.” In: *British Journal of Surgery* 103.3 (Feb. 2016), pp. 300–306.
- [321] E Matt Ritter and Daniel J Scott. “Design of a proficiency-based skills training curriculum for the fundamentals of laparoscopic surgery.” In: *Surgical innovation* 14.2 (June 2007), pp. 107–112.
- [322] Gabriele Wulf and Charles H Shea. “Principles derived from the study of simple skills do not generalize to complex skill learning”. In: *Psychonomic Bulletin & Review* 9.2 (June 2002), pp. 185–211.
- [323] P J Wraighte, D P Forward, and P Manning. “The Impact of the European Working Time Regulations on Orthopaedic Trainee Operative Experience”. In: *Bulletin of The Royal College of Surgeons of England* 94.5 (May 2012), pp. 156–158.

- [324] G Morrow et al. *The Impact of the Working Time Regulations on medical education and training: Final report on primary research*. Tech. rep. Durham University, Aug. 2012.
- [325] Andrew R L Medford. “Post-certificate of completion of training fellowships.” In: *Clinical Medicine (London, England)* 9.5 (Oct. 2009), pp. 441–443.
- [326] Timothy Barlow, David Barlow, and Alison Edwards. “Consultant contracts in England: how many SPAs do you have?” In: *BMJ* (June 2012), e3594.
- [327] A Hyltander et al. “The transfer of basic skills learned in a laparoscopic simulator to the operating room.” In: *Surgical Endoscopy* 16.9 (Sept. 2002), pp. 1324–1328.
- [328] Peter J M Ceponis et al. “A randomized pilot validation of educational measures in teaching shoulder arthroscopy to surgical residents.” In: *Canadian Journal of Surgery. Journal Canadien de Chirurgie* 50.5 (Oct. 2007), pp. 387–393.
- [329] Chris C Cychosz et al. “Fundamentals of Arthroscopic Surgery Training Program Improves Knee Arthroscopy Simulator Performance in Arthroscopic Trainees”. In: *Arthroscopy: The Journal of Arthroscopic & Related Surgery* (Feb. 2018).
- [330] J Shah and A Darzi. “Surgical skills assessment: an ongoing debate.” In: *BJU International* 88.7 (Nov. 2001), pp. 655–660.
- [331] J A Martin et al. “Objective structured assessment of technical skill (OSATS) for surgical residents.” In: *British Journal of Surgery* 84.2 (Feb. 1997), pp. 273–278.
- [332] Daniel Banaszek et al. “Virtual Reality Compared with Bench-Top Simulation in the Acquisition of Arthroscopic Skill: A Randomized Controlled Trial.” In: *The Journal of Bone and Joint Surgery. American volume* 99.7 (Apr. 2017), e34.
- [333] Rajesh Aggarwal and Teodor Grantcharov. “Prospective, Randomized Assessment of Transfer of Training (ToT) and Transfer Effectiveness Ratio (TER) of Virtual Reality Simulation Training for Laparoscopic Skill Acquisition”. In: *Annals of Surgery* 262.1 (July 2015), e33–e34.
- [334] Vanessa N Palter and Teodor P Grantcharov. “Individualized Deliberate Practice on a Virtual Reality Simulator Improves Technical Performance of Surgical Novices in the Operating Room”. In: *Annals of Surgery* 259.3 (Mar. 2014), pp. 443–448.
- [335] Faisal Aziz. “Vascular Surgery Trainees Still Need to Learn How to Sew: Importance of Learning Surgical Techniques in the Era of Endovascular Surgery”. In: *Frontiers in Surgery* 2 (May 2015), pp. 54–2.
- [336] Andrew J Carr. “Arthroscopic surgery for degenerative knee.” In: *BMJ* 350 (2015), h2983.
- [337] A F Anderson et al. “Interobserver Reliability of the International Society of Arthroscopy, Knee Surgery and Orthopaedic Sports Medicine (ISAKOS) Classification of Meniscal Tears”. In: *The American Journal of Sports Medicine* 39.5 (May 2011), pp. 926–932.
- [338] Vivek Datta et al. “The relationship between motion analysis and surgical technical assessments.” In: *The American Journal of Surgery* 184.1 (July 2002), pp. 70–73.
- [339] Toby Hillman. *Medicine – a team sport*. Jan. 2014. URL: <http://blogs.bmj.com/pmj/2014/01/13/medicine-a-team-sport/>.

- [340] Karolina Wartolowska et al. "Use of placebo controls in the evaluation of surgery: systematic review." In: *BMJ* 348.may21 2 (May 2014), g3253–g3253.
- [341] Abtin Alvand. "Improving surgical learning and performance at unicompartmental knee arthroplasty". PhD thesis. Oct. 2014.
- [342] Richard J Parent et al. "Early, intermediate, and late effects of a surgical skills "boot camp" on an objective structured assessment of technical skills: a randomized controlled study." In: *Journal of the American College of Surgeons* 210.6 (June 2010), pp. 984–989.
- [343] Marilyn Strathern. "Improving ratings': audit in the British University system". In: *European Review* 5.3 (1997), pp. 305–321.
- [344] Maheen M Adamson et al. "The impact of brain size on pilot performance varies with aviation training and years of education." In: *Journal of the International Neuropsychological Society : JINS* 16.3 (May 2010), pp. 412–423.
- [345] Eleanor A Maguire et al. "Navigation-related structural change in the hippocampi of taxi drivers". In: *Proceedings of the National Academy of Sciences* 97.8 (Apr. 2000), pp. 4398–4403.
- [346] Vanessa Sluming et al. "Voxel-Based Morphometry Reveals Increased Gray Matter Density in Broca's Area in Male Symphony Orchestra Musicians". In: *NeuroImage* 17.3 (Nov. 2002), pp. 1613–1622.
- [347] Beatriz Calvo-Merino et al. "Seeing or Doing? Influence of Visual and Motor Familiarity in Action Observation". In: *Current Biology* 16.19 (Oct. 2006), pp. 1905–1910.
- [348] Kenoki Ohuchida et al. "The frontal cortex is activated during learning of endoscopic procedures". In: *Surgical Endoscopy* 23.10 (Jan. 2009), pp. 2296–2301.
- [349] Marie C Morris et al. "Assessment of competence in surgical skills using functional magnetic resonance imaging: a feasibility study." In: *Journal of Surgical Education* 72.2 (Mar. 2015), pp. 198–204.
- [350] Kyle R Wanzel et al. "Visual-spatial ability and fMRI cortical activation in surgery residents". In: *The American Journal of Surgery* 193.4 (Apr. 2007), pp. 507–510.
- [351] Hemel Narendra Modi et al. "A decade of imaging surgeons' brain function (part II): A systematic review of applications for technical and nontechnical skills assessment." In: *Surgery* 162.5 (Nov. 2017), pp. 1130–1139.
- [352] Parisa Bahrami et al. "Functional MRI-compatible laparoscopic surgery training simulator." In: *Magnetic Resonance in Medicine* 65.3 (Oct. 2010), pp. 873–881.