

**THE VALIDITY, RELIABILITY AND  
RESPONSIVENESS OF PROCEDURE  
BASED ASSESSMENT IN  
SIMULATED VASCULAR  
PROCEDURES**

---

**One Volume**

**By Miss Lucy Anne Green BSc, MBChB, MRCS,  
MClined**

**Being a thesis submitted for a Doctorate of  
Medicine (MD)**

**From the University of Hull and University of  
York**

**Hull York Medical School**

**Submitted September 2014**

**Word Count: 46,115**

## **Abstract**

### **Introduction**

Procedure based assessment (PBA) has been shown to be valid and reliable in the workplace however whether this translates to the simulation setting has not yet been demonstrated. Therefore the aim of this thesis is to demonstrate that PBA is a valid, reliable and responsive assessment tool in simulated vascular procedures.

### **Methods**

Three experiments based on simulated vascular operations were designed to explore the validity, reliability and responsiveness of PBA utilising 3 commonly performed vascular procedures. The global and task specific checklist (GTSC) and global summary score (GSS) of a modified PBA were analysed separately. Validity was determined by correlating performance with prior operative experience (number of operations previously observed and performed) and stage in surgical training. Reliability and responsiveness was determined by use of multiple raters and assessing change in performance over time.

### **Results**

The modified PBA was found to be a valid assessment method based on number of operations previously performed ( $r=0.446$   $p=0.029$  for the GTSC and  $r=0.553$   $p=0.005$  for the GSS) but not for operations previously observed. Only the PBA GTSC was valid for stage of surgical training ( $r=0.588$   $p=0.002$ ). The modified PBA demonstrated good inter-rater reliability ( $r=0.665$   $p=0.005$  for the GTSC and  $r=0.843$   $p>0.001$  for the GSS) during simulated vascular procedures. Intra-rater reliability was not demonstrated. The PBA GSS was found to be responsive to improved performance (WSR  $p<0.001$ ) but the PBA GTSC was not (WSR  $p=0.104$ ).

## **Conclusion**

The modified PBA is a valid assessment of surgical skill when correlated with previous operative performance. Observation alone appears to contribute little to assessment outcomes. Performance is index specific and not fully dependant on training level. PBA has only partial reliability in simulated vascular procedure due to the lack of intra-rater reliability. PBA was responsive to practice effect suggesting it could be useful to monitor trainee performance in simulation. PBA potentially has a role in simulation assessment but did not demonstrate sufficient reliability for high stakes examination.

## Contents

Abstract.....	2
List of tables and illustrations.....	8
List of accompanying material.....	10
Preface.....	11
Acknowledgements.....	13
Authors declaration.....	14
Chapter 1: Introduction.....	15
1.1 Background.....	15
1.2 Study Aims.....	19
1.3 Validity, Reliability and Responsiveness.....	20
1.2.1 Validity.....	23
1.2.2 Reliability.....	29
1.2.3 Responsiveness.....	31
1.3 Assessment methods in surgical technical skills and simulation.....	33
1.3.1 Background.....	33
1.3.2 Rater free assessment methods.....	36
1.3.3 Assessment Methods Requiring a Rater.....	42
Kopta’s Observation Guide.....	44
Experimental surgical assessments tools.....	44
Directly observation of procedural skill (DOPS).....	45
Objective Assessment of Technical Skills (OSATS).....	46
Procedure Based Assessment (PBA).....	52
1.4 The case for Simulation: An Overview of the Evidence.....	64
1.4.1 Bench Models.....	68
1.4.2 Endovascular Simulation.....	78
1.4.3 Video Analysis.....	85
1.5 Chapter Summary.....	90
Chapter 2: Materials and Methods.....	92
2.1 Chapter Overview.....	92
2.2 Data Collection Methods.....	92

2.3 Statistical analysis .....	95
2.4 Inclusion and exclusion criteria .....	97
2.5. Data Security and Confidentially .....	98
2.5 Experiment 1- Validity and Inter- Rater Reliability of Procedure Based Assessment in Simulated Vascular Procedures .....	99
2.5.1 Study Setting for Experiment 1 .....	99
2.5.2 Study participants.....	99
2.5.3 Study Assessors .....	99
2.5.4 Simulated Procedures and Materials.....	100
2.5.5 Study Protocol .....	103
2.5.6 Good Clinical Practice and Ethical Considerations .....	103
2.6 Experiment 2- Construct validity of the ANGIOMentor endovascular simulator and inter- rater reliability and responsiveness of Procedure Based Assessment in simulated superficial femoral artery angioplasty .....	105
2.6.1 Study setting for experiment 2.....	105
2.6.2. Study participants.....	105
2.6.3 Study Assessors .....	106
2.6.4 Materials and Procedures Assessed .....	106
2.6.5 Study protocol .....	110
2.6.6 Good clinical practice and ethical considerations .....	111
2.7 Experiment 3- Intra-rater Reliability of Procedure Based Assessment in simulated vascular procedures. ....	112
2.7.1. Study Setting for experiment 3 .....	112
2.7.2 Study participants.....	112
2.7.3 Study assessors.....	112
2.7.4 Materials and procedures assessed .....	112
2.7.5 Study protocol .....	113
2.7.6 Good clinical Practice and ethical considerations .....	114
Chapter 3- Results.....	115
3.1 Experiment 1- Validity and Reliability of Procedure based assessment in simulated vascular procedures .....	115
3.1.1 Participant recruitment and baseline characteristics.....	115
3.1.2 Validity of Global and Task Specific Checklist (GTSC) and Global Summary Score (GSS) of Procedure Based Assessment for number of procedures previously observed .....	118

3.1.3 Validity of Global and Task Specific Checklist (GTSC) and Global summary score (GSS) of Procedure Based Assessment for number of procedures previously performed.....	120
3.1.4 Validity of the Global and Specific Checklist and Global Summary Score of Procedure Based Assessment in relationship to training grade .....	122
3.1.5 Validity of Objective Structured Assessment of Technical Skills in assessment of previous observation, performance and training grade.....	124
3.1.6 Concurrent validity between Objective Assessment of Technical Skills and the Procedure Based Assessment .....	126
3.2 Experiment 2- Inter-rater Reliability and Responsiveness of Procedure Based assessment in simulated vascular procedures.....	128
3.2.1 Part a) Determining the validity of the ANGIOMentor Simulator .....	128
3.2.2 Part B) Participant recruitment and baseline data for Inter-rater reliability and responsiveness .....	133
3.3 Inter-rater reliability.....	136
3.3.1 Inter-rater reliability of the Procedure based assessment global and task specific checklist (GTSC) for simulated vascular procedures.....	136
3.3.2 Inter-rater reliability of the Procedure Based Assessment Global Summary Score (GSS) simulated vascular procedures .....	137
3.4 Responsiveness of Procedure Based Assessment in simulated vascular procedures .....	138
3.4.1 Responsiveness - Cohort data .....	141
3.5 Experiment 3- Intra-rater reliability of PBA.....	142
3.5.1 Participant recruitment and base line data .....	142
3.5.2 Results for Intra Rater Reliability.....	142
3.6 Results Summary .....	145
Chapter 4- Discussion.....	147
4.1 Chapter Overview.....	147
4.2.1 Validity .....	148
4.2.2 Reliability.....	151
Inter-rater reliability.....	151
Intra- rater reliability .....	153
4.2.3 Responsiveness .....	155
4.3 Limitations of this research.....	157
4.4 Implications of this research .....	159

4.5 Future directions .....	160
4.6 Dissemination of work.....	166
4.7 Conclusion.....	166
References.....	168
Appendix 1-ICEPS.....	180
Appendix 2- DOPS .....	181
Appendix 3-An example of the original OSATS tool and generic OSATS.....	182
Appendix 4- PBA .....	185
Appendix 5 – Modified PBA for Sapheno-femoral junction ligation, proximal aortic anastomosis and superficial femoral artery angioplasty.....	188
Appendix 6- Modified Generic OSATS.....	193
Appendix 7 Formative Supervised Surgical Assessment .....	194
Glossary of terms and definitions .....	196

## List of tables and illustrations

Table 1 Summary of the main assessment devices which provide rater free feedback during assessment of surgical psychomotor skill.....	41
Table 2 summary of surgical psychomotor assessments which require human rater input .....	43
Table 3 Summary of the main historical studies and relevant vascular surgical studies utilising Objective Structured Assessment of Technical Skills.....	50
Table 4 Summary of the relevant studies historic studies utilising Procedure Based Assessment.....	58
Table 5 Example of the output from a “d study” on reliability of Procedure based assessment in the workplace.....	62
Table 6 Summary of the main vascular simulation studies involving bench models and similar techniques .....	75
Table 7 Summary of the main endovascular simulation studies illustrating the simulator type and procedure investigated .....	84
Table 8 Summary of the main vascular simulation studies utilising video assessment method .....	89
Table 9 Distribution of assessors and trainees performing sapheno-femoral junction ligation and proximal aortic anastomosis.....	117
Table 10 Original data from all paired PBA assessments from Experiment 1 and Experiment 2.....	135
Table 11 Primary data from experiment 2 illustrating individual participant performance between the 5 <sup>th</sup> and 10 <sup>th</sup> simulator run on the ANGIOMentor simulator .....	139
Table 12 All matched pairs of PBA on the ANGIOMentor (1st and 2nd assessed runs) .....	140
Table 13 Primary data from experiment 3 illustrating the scores provided by individual raters between their primary and secondary assessment of the simulated sapheno-femoral junction ligation. ....	143
Figure 1 Schematic diagram illustrating how the differing definitions of validity are related to construct validity .....	28
Figure 2 Readout from the ANGIOMentor endovascular simulator for superficial femoral artery angioplasty .....	38
Figure 3 An example of a low fidelity vascular surgical training jig .....	65
Figure 4 Illustration of the primary incision during the saphenofemoral junction ligation.....	101

Figure 5. Illustration of ligated tributaries required during the dissection of the sapheno-femoral junction ligation.....	101
Figure 6 Illustration of the aortic aneurysm simulator .....	102
Figure 7 Illustration of the completed proximal aortic anastomosis.....	102
Figure 8 The ANGIOMentor simulator illustrating the two display readout and joystick controls for the x-ray tube and table.....	108
Figure 9 Illustration of the catheters and wires used to insert into the simulator to perform the interventional procedure of choice.....	108
Figure 10 An example of the fluoroscopy readout after contrast injection of the left femoral access point .....	109
Figure 11 illustrating the simulators ability to rotate the lesion of interest in 3 D to aid teaching and catheter and wire positioning .....	109
Figure 12 Scatter graph illustrating the distribution of PBA GTSC scores and Spearman's correlation coefficient with number of operations previously observed/assisted with .....	119
Figure 13 Scatter graph illustrating the distribution of PBA GSS and Spearman's correlation co-efficient with number of operations previously observed/ assisted with .....	119
Figure 14 Scatter graph illustrating the distribution of PBA GTSC scores and Spearman's correlation co-efficient with number of operations previously performed.....	121
Figure 15 Scatter graph illustrating the distribution of PBA GSS and Spearman's correlation co-efficient with number of operations previously performed.....	121
Figure 16 Scatter graph illustrating the distribution of PBA GTSC and Spearman's correlation co-efficient with year in surgical training.....	123
Figure 17 Scatter graph illustrating the distribution of PBA GSS and Spearman's correlation co-efficient with year of surgical training.....	123
Figure 18 Scatter graph illustrating the distribution of OSATS scores and Spearman's correlation co-efficient with number of operations previously observed.....	124
Figure 19 Scatter graph illustrating the distribution of OSATS scores and Spearman's correlation co-efficient with number of operations previously performed.....	125
Figure 20 Scatter graph illustrating the distribution of OSATS scores and correlation co-efficient with year in surgical training.....	125
Figure 21 Bland-Altman Plot describing the relationship and level of agreement between OSATS and the ATIS PBA Scores. ....	126
Figure 22 Bland-Altman Plot illustrating the relationship and level of agreement between OSATS and the PBA GSS.....	127
Figure 23 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for total procedure time.....	130
Figure 24 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for total fluoroscopy time.....	130

Figure 25 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for time to position sheath for intervention .....	131
Figure 26 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for residual stenosis of the SFA lesion after angioplasty .....	131
Figure 27 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for contrast volume used .....	132
Figure 28 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for advancement of the guide or sheath without a leading wire .....	132
Figure 29 Scatter graph illustrating the distribution of PBA GTSC scores and Spearman's correlation co-efficient between 1 <sup>st</sup> and 2 <sup>nd</sup> raters .....	136
Figure 30 Scatter graph illustrating the distribution of PBA GSS and correlation co-efficient between 1 <sup>st</sup> and 2 <sup>nd</sup> raters .....	137
Figure 31 Box and whisker graph illustrating the cohort scores for the PBA GTSC between the 5 <sup>th</sup> and 10 <sup>th</sup> run on the ANGIOMentor simulator .....	141
Figure 32 Box and whisker graph illustrating the cohort score for the PBA GSS between the 5 <sup>th</sup> and 10 <sup>th</sup> run on the ANGIOMentor simulator .....	141
Figure 33 Scatter graph illustrating Spearman's correlation between the PBA GTSC scores between the primary and secondary assessment of the simulated sapheno-femoral junction ligation.....	144
Figure 34 Scatter graph illustration the Spearman's correlation for the PBA GSS between the primary and secondary assessment of the simulated sapheno-femoral junction ligation.....	144

## List of accompanying material

Video of Sapheno-femoral Junction Ligation(DVD).....	Back Cover
------------------------------------------------------	------------

## **Preface**

This work arose due to the many frustrations I had as a surgical trainee undertaking and defining the purpose of Procedure Based Assessment (PBA) within the context of my surgical training. Like many trainees and trainers I often saw PBA as a tick box exercise that contributed little to my training.

Undertaking a period of research and supported by studies for a Masters in Clinical Education I began to see the role of PBA within the surgical curriculum. While PBA was designed to be an “assessment for learning” it has also become clear that it is an assessment designed to provide “evidence of learning” certainly this has become the case during many trainees yearly appraisal process.

Given the weight now placed on work placed based assessments in ensuring trainees can progress satisfactorily in their training I felt that evidence should be sought to establish that PBA represented a meaningful assessment of trainee performance i.e. was it a valid, reliable and responsive tool to monitor trainee performance and progression? Simulated procedures are increasingly being incorporated into surgical training and seemed an ideal media that were readily available to explore the psychometric properties of PBA and so this thesis was born.

I was unaware at the beginning of this research that there were other surgical assessment tools such as the Objective Structured Assessment of Technical skills (OSATS) or Imperial College Evaluation of Procedural Technical Skills (ICEPS) which may have been better suited to evaluating simulated assessments; however since PBA was the main WBA that was in common use in the UK at the time, I persevered.

The Yorkshire Deanery had also recently purchased a number of endovascular simulators one of which was sitting underutilised in the clinical skills centre at Hull Royal Infirmary. This also seemed a ideal opportunity to see if PBA could be applied to assessment of endovascular procedures as well as open vascular simulation.

Since undertaking this research PBA has been studied in the workplace and that research has answered some of these questions which are discussed further in this thesis. However, hopefully this study will add to the evidence base of PBA and potentially define a role for PBA in surgical simulation training.

The research undertaken in this thesis was performed between September 2010 and September 2012. Application for registration of the degree of MD was made in September 2011 and the final thesis was submitted after a write up period of 1 year from September 2013 to September 2014.

## **Acknowledgements**

I would like to dedicate this thesis to my son Neal Gray who came into this world on the 29<sup>th</sup> of July 2014 and his father, my partner Stuart. While this thesis represents a significant academic achievement Neal is undoubtedly my finest work and Stuart has provided love and support (not to mention endless cups of tea!) and put up with my occasional irritable outbursts while I have endeavoured to finish this work.

I would also like to thank Professor I.C Chetter, my academic supervisor, for encouraging me to undertake this M.D, his guidance in helping me select the topic and direction of research and advice on thesis structure and keeping things simple! I will always be grateful for the opportunity to enter the world of research and broaden my horizons.

In particular I would also like to thank Mr Brian Johnson from my thesis advisory panel who has given his time, knowledge and enthusiasm without recompense. He has provided guidance, support and critical feedback and his invaluable proof reading skills.

I would also like to thank Victoria Allgar, the University of Hull statistician who provided statistical advice and pointed me in the right direction in analysing my data.

I would like to thank all those individuals in Hull and East Yorkshire NHS Trust and the Royal College of Surgeons, in particular Vishy Mahadevan; Professor of Anatomy, who were involved in this study and gave their time and labour and without whom this thesis would not of been possible.

Finally, to my parents' thank you for believing in me, your advice, support and constant nagging to get this thesis finished. I couldn't have done it without you!

## **Authors declaration**

"I confirm that this work is original and that if any passage(s) or diagram(s) have been copied from academic papers, books, the internet or any other sources these are clearly identified by the use of quotation marks and the reference(s) is fully cited. I certify that, other than where indicated, this is my own work and does not breach the regulations of HYMS, the University of Hull or the University of York regarding plagiarism or academic conduct in examination. I have read the HYMS code of Practice on Academic Misconduct, and state that this piece of work is my own and does not contain any unacknowledged work from any other sources'. I confirm that any participant information obtained to produce this piece of work had been appropriately anonymised"

## **Chapter 1: Introduction**

### **1.1 Background**

Assessment of surgical and technical skills in trainees has until the last three decades been a poorly researched area of surgical curriculum delivery and education, hampered in large part by the ever changing components of curriculum and desirable characteristics which we wish to measure in the surgical trainee(1) . It has been an evolutionary process, which has been driven by a number of influences.

Early assessment processes lead by individual centres of learning focused on structure and process based assessment of students where knowledge acquisition was the focus of the curricula and the ability to remember and regurgitate facts on demand was the ethos of such assessment(2). Models of surgical education were until relatively recently based upon an apprenticeship acquired through long hours of observation and service provision(3) . Progression through the disciplines of medicine, surgery and many other specialities is still governed through formal examination administered by the Royal Colleges where knowledge is tested in a formal, norm referenced, pass or fail manner based around knowledge acquisition. But they too, are beginning to be influenced by the changes in surgical education and research and the requirement for comprehensive procedural skill assessment as well as knowledge. This is evidenced by the introduction of Objective Structured Clinical Examination (OSCE) into the surgical examination process at membership level. However this is still not yet part of the final Certificate of Completion of Training (CCT) which is required before trainees can take up a consultant post.

Prior to 2007 when workplace based assessments were introduced through implementation of the Intercollegiate Surgical Curriculum Project (ISCP)(4)

the surgical assessment of the psychomotor skills necessary to perform as a safe independent practitioner was inferred by time served, log books and yearly appraisal processes that were dependant on subjective supervisor reports rather than an objective appraisal of surgical skill(5, 6).

Over time the discipline of surgical education has evolved to become a research area in its own right and has had a growing impact on how surgical curricula are designed, implemented and indeed measured(7). The rise of outcome and objective based curricula, the earliest example of which is described by Ralph W Tyler in the 1950's which advocated a broad view of learning objectives(8), has had large impact on surgical curricula which is evidenced by the present incarnation of the ISCP where learning outcomes are behaviourally defined. This approach has been advocated by the General Medical Council(9) and been incorporated in to the undergraduate and post graduate curriculum in a wholesale manner and has had a heavy impact on how workplace based assessments are structured and implemented(10).

The changes in surgical assessment have also been politically driven and influenced. There is a growing demand by the general public to see a more rigorous evaluation of practicing surgeons and doctors. Notably in the wake of the Bristol Heart and Shipman inquiries where the behaviour, probity and technical ability of practicing doctors was questioned(11, 12).

Together these drivers had brought about the development and evolution of the work placed based assessments (WBA) and in particular the surgical procedural assessment known as Procedure Based Assessment (PBA) and, in the coming years, for formal revalidation processes that will objectively look at medical practitioner's performance in the workplace(13).

However Procedure Based Assessment was introduced into practice as an aid for trainees to receive feedback on their performance in a formative manner. But there has been a growing emphasis on using PBA as

summative assessment tool with increasing numbers being required at each Annual Review of Competency Progression (ARCP) in order to allow progression through to the next stage of training. Some may argue this negates the very purpose of its inception which is primarily as an aid to learning(14).

This change of purpose means that PBA should be subject to scrutiny as a tool of assessment and yet we have little knowledge as yet on whether it is valid, responsive and a reliable assessment tool in assessment of and progression of psychomotor surgical skill. We also do not know if using PBA can result in a superior product, the practising surgeon. Indeed except for the WBA known as multisource feedback (MSF), which has been shown to confer a change in behaviour (15) there is little other than inference that it is a beneficial assessment.

The increasing emphasis on safety and scrutiny of surgical performance in national audits, combined with decreasing exposure to theatre time due to the constraints of the European Working Time Regulation(16), mean that surgical programme directors and trainers are looking for additional ways to train the next generation of surgeons more efficiently(17). The competency defined curriculum as opposed to a time limited programme which incorporates PBA may provide one answer. However this means that such assessment methods have to be rigorously assessed and validated in order to ensure that they confidently reflect the trainees performance if we are to rely on them to inform progression. An alternative, which may provide part of the solution, is to assess competencies using simulated procedures (18-20). This approach circumvents some of the issues of obtaining access to theatre lists; which has been problematic, can provide a controlled environment and reduce some of the variables that can affect trainee performance. This may involve difficulty of the procedure, fatigue and the high stakes element which is present when assessing trainees

during live operations(21). Increasingly complex high fidelity simulators are being designed to replicate part or all of many index procedures available in the ISCP (22-26). There is a growing body of evidence that simulation can shorten learning curves and increase trainee confidence. It can also provide a learning experience equal to that provided by operating in theatre on patients, but without the risk of complications(27-31).

Controlled deliberate practice develops expertise(32, 33) and when combined with experienced feedback from a senior clinician can provide a powerful adjunct to traditional training methods(34-38). But in the arena of surgical practice we have in the past poorly evaluated simulated surgical skills courses where attendance and participation have been mandatory but where little effort has been made to formally assess and benchmark standards and this is the very essence of PBA as it sits at the moment within the ISCP and surgical practice: it is a bench marking tool for assessing competency in a particular procedure but there is no evidence as yet to assume it can infer competence in other areas of surgical practice or procedures.

Combining the use of PBA in assessing surgical simulation may allow all of the potential of PBA to be fulfilled, as a tool to provide feedback, to monitor progression, to bench mark standards of surgical skill in a controlled environment and if necessary to provide evidence of technical competence in the revalidation process(21, 39).

At present the evidence base for PBA is scanty and this will be explored in the following chapters. The purpose of this thesis and in the experiments designed within this study will be determine if PBA has a role in the simulated environment and if it is a valid, reliable and responsive tool to assess simulated vascular procedures.

## 1.2 Study Aims

The main study aim of this thesis will be to determine if Procedure Based Assessment has the potential to be a valid, reliable and responsive tool in assessing vascular simulation technologies. The following study aims will be explored:

- Is PBA a valid measure of surgical performance as measured by previous surgical experience; as determined by number of operations previously performed and observed and training level in vascular simulated procedures?
- Does PBA possess concurrent validity with a previously validated assessment tool in surgical simulation; Objective Structured Assessment of Technical Skills?
- Is PBA a reliable measure of surgical performance as determined by inter-rater and intra-rater reliability observing simulated vascular surgical procedures?
- Is PBA a responsive tool in measuring surgical performance as determined by observation of simulated endovascular surgical procedures?

### **1.3 Validity, Reliability and Responsiveness**

The evaluation of assessment tools such as PBA which measure observable behaviours such as surgical skills is known as psychometrics(40). This is a discipline which involves the development and refinement of psychological measurement, typically involving measuring behaviour, knowledge, attitude and educational assessment tools. Sir Francis Galton is credited with the creation of the discipline of psychometrics in his research into the area of how humans differ from each other, initially in their levels of intelligence, strength or motor function and was informed by ideas from his cousin Sir Charles Darwin and his seminal work the Origin of the Species(41). Galton was credited with developing regression and correlation statistical analysis and was particularly keen on measuring the measurable and felt something was only of scientific importance if it could be quantified. At similar time a number of German psychophysicist such as Gustav Fechner and Wilhelm Wundt were experimenting in measuring stimulus and response behaviour and are responsible for some of the first descriptions of experimental psychology. In conjunction with the early statisticians at that time including Spearman and Galton a discipline arose that was able formulate models and methods to measure empirically collected psychological data that then allowed the researcher to draw behavioural inferences. In 1935 the Psychometric Society was founded which is concerned with three main areas of research; psychological scaling, educational and psychological measurement and factor analysis(42).

Applying psychometric evaluation to educational assessment tools such as Workplace Based Assessments can be facilitated by exploring the Utility formula(43). The Utility formula defines the desirable characteristics that

potentially can be measured to determine if an educational assessment has the potential to be worthwhile.

**Educational impact x reliability x validity x cost effectiveness x  
acceptability(44)**

Clearly the educational impact of educational assessment tools such as WBA are extremely difficult to measure and quantify as the effect of such assessments cannot be fully evaluated until trainees have been through the defined curriculum and surgical system and the effect of this training method evaluated, an outcome in itself that is not easily quantified or measured. Likewise cost effectiveness and acceptability are also heavily linked to whether a tool can demonstrate an educational impact on the trainee. As a poor educational impact would reduce any intervention's cost effectiveness and ultimately acceptability to the trainee, trainer, curriculum designer and the general public who have a vested interest in ensuring money is well spent and surgeons and doctors are trained to the highest possible standard.

A surrogate for educational impact can be considered the sensitivity of a WBA to detect change over time; a property known as responsiveness (45, 46). This is particularly important in simulation training where students appreciate being able to monitor their progress and have clearly defined benchmarks and goals to work towards(47).

Determining validity and reliability and indeed responsiveness of an assessment tool are early key fundamental steps in determining the usefulness of any assessment tool(48) and one of the simplest ways to measure educational assessment models in the discipline of psychometrics is known as classical test theory. This theory assumes that raw scores such as assessment scores are measurements and can be assigned a value according to predefined rules. The research then focuses on determining if the scores are associated with underlying variables or associations. The

basic premise is that a total test score is made up of multiple items which contain a true score and random error component, both of which can be measured(49). But there are limitations to classical test theory as the results developed from the data set are dependent on the sample being tested and so can limit the generalizability of the data unless the sample that is being used is highly representative of the study population.

The utility formula provides a useful framework to assess the psychometric properties of PBA in a quantitative manner. Particularly so in a simulation setting where certain behaviours are being assessed. Indeed the PBA form itself is a list of statements of behaviours which are normally assessed in a binary fashion. Certainly validity, reliability and responsiveness are properties that can be mathematically measured and processed to test a statistical hypothesis. Factors such as cost effectiveness, acceptability and feasibility which are often quoted in addition to Cees Van der Vleutens' original utility formula(50) are perhaps best explored with qualitative research methods where a framework such as SWOT analysis can be utilised to assess the strengths and weaknesses of an assessment process rather than the tool itself. While the properties of acceptability, feasibility and cost effectiveness are clearly important to the functionality of the assessment in the real world if the assessment does not at least possess a degree of validity and reliability then it is not representative of the curriculum being assessed nor can it be relied upon for examination and summative purposes.

This highlights a well known dichotomy as to what constitutes a perfect assessment tool; while all the factors highlighted in the utility formula are clearly important whether the tool is to be used as a summative assessment instrument or formative purposes will dictate how important each of these factors should be. In order for a behavioural assessment tool to be highly representative of a behavioural curriculum objectives it should

possess a high degree of construct validity i.e. measures and differentiates between the psychometric skills it has been designed to measure. In contrast, reliability is not relevant to the individual learner, where feedback and validity are the most important. Reliability is important for summative and pass/fail purposes, indeed an assessment may be highly reliable but this may be at the cost of validity and the ability of an assessment tool to provide feedback and a useful learning experience(50).

### **1.2.1 Validity**

Validity in psychometric measurement encompasses a range of diverse definitions that reflect subtly different ways of measuring validity. Review of educational literature often highlights this fact and readers are faced with a bewildering array of definitions such as face, content, concurrent, convergent, divergent and construct validity(40, 45, 48). This can make interpretation of educational literature on assessment instruments difficult and comparison between studies nearly impossible as frequently authors present data on only one element of validity which can cast doubt on the validity of that instrument as a whole(51).

Unification of the concept of validity has been explored. Lee J. Cronbach and Paul E. Meehl in 1955(cited by Colliver 2012(52)) presented a paper in which they define predictive, concurrent, content and construct validity. They state the first two types of validity are what are known as criterion orientated validation procedures or where evidence is sought of a relationship between the attributes in a measurement tool and the variable under study(52). This is contrasted with DeVon et al(48) who suggest that construct validity is an overarching concept, which is reflective of more current thinking, where concurrent, predictive, convergent and discriminant validity are forms of criterion related validity as opposed to

face and content validity which are forms of translational validity: a term used to imply a tool is a translation of the construct under study. The relationships between these types of validity are illustrated in figure 1 (page 28).

However the consensus that emerges from the literature is that construct validity is desirable when other forms of validity are deemed to be insufficient and that it can be defined by the statement “does the test or evaluation measure what it purports to measure” according to Cronbach and Meehl (cited in Moss 1992(53)) suggest that for such construct validity to be scientifically correct it had to demonstrate observable or statistical laws which tie observations to one another and also central to this concept was the directionality of the hypothesis under study. Since that time others have argued that construct validity should also consider the social consequences of how the data was used (53-56) which depending on the type of assessment administered may have ethical considerations but certainly links in with the broad outline of acceptability within the Utility formula.

In the context of this study the assessment tool being studied is measuring desirable behaviours in a surgical trainee rather than knowledge. Content and face validity of PBA has already been established, and so the concept of construct validity as defined by DeVon(48), an overarching concept, best fits the investigation of a tool that has been designed to determine differences in surgical performance and progression.

Nested within the concept of construct validity are the many lesser types of validity as outlined in figure 1 (p28), only some of which would be applicable to PBA in this instance but none which would describe the relationship between PBA scores and surgical performance which is one of the aims of this thesis. These lesser types of validity can contribute but not

necessarily confer validity of an assessment method or tool and can be defined as the following:

- Face validity is conferred to an assessment when the tool or method appears to represent the construct under study. An example of this would be an assessment tool such as an itemised check list which appears to represent each step of a procedure that needs to be evaluated. Classified as translational it appears to reflect or translate from the construct under study. It is considered less than robust due to the lack of statistical or observational rigour applied to its inception(45).
- Content validity may appear at first glance to be very similar to face validity; the content of an assessment method or tool reflecting the assessment itself. However in this instance the content of the assessment has been derived from an iterative process often including a panel of experts who evaluate each question or item within an assessment to determine what should be included. Items or questions are often derived from a pool of information gleaned from the literature, expert opinion or qualitative research. Statistical rigour can be applied in this instance and one method that illustrates this process well is where items can be rated by each expert to calculate a Content Validity Ratio (CVR) which is then compared to a table of minimum values with a significance of .05. The remaining item ratios are then converted to a mean to give a Content Validity Index (CVI) for that particular tool or assessment method(57).
- Concurrent validity is conferred when a new instrument which is being validated is compared and statistically correlated with a pre-existing, previously validated instrument which has been designed to measure the same construct or concept under study and is

considered to be the gold standard in measuring that particular construct(45, 48). An example would be two different questionnaires that have both been designed to measure quality of life in patients with vascular disease. Key to concurrent validity is that the instruments are evaluated together at the same time on the same population and should give similar scores that correlate to a significant degree.

- Convergent validity may sound very similar to concurrent validity in the fact that a researcher would expect two similar tools to converge or correspond. But instead would be measuring similar or related constructs not the same construct, such as children's pain scores after surgery and the need for analgesia post operatively. The questionnaires or instrument may not have been previously validated, nor represent the gold standard but should theoretically give similar scores or correlate well(45, 48).
- Discriminant validity is the opposite of convergent validity. Here the researcher would expect scores from different instruments that measure opposite but related concepts such as health and well being and pain scores for joint disease to correlate poorly or diverge. This would reassure the researcher that the tool under study has the capacity to discriminate between patients that are well and those are suffering with pain and disability from joint disease(48).
- Predictive validity is conferred when a new instrument is administered within a particular cohort, the results of which are then later compared to a known gold standard test(45) or even the same instrument(48) . If the scores from the first instrument correlate highly with the scores from the second test within the

same cohort, the test or instrument is said to have predictive validity.

What becomes evident though the literature is that validity is a concept rather than an absolute measurement. It is conferred by demonstrating a relationship between the instruments scores and interpretation or hypothesis of the researcher and not the tool itself(40). There is no single agreed statistical measurement which, as discussed previously, reflects the various forms of validity. Papers variously report different statistical indices, such as correlation coefficients like Spearman's and Pearson's  $r$ . Mann-Whitney U and Kruskal- Wallis and ANOVA have also been used to describe or prove directionality of a hypothesis regarding validity of assessment tools, particularly in surgical education(51). Validity is often claimed when these results reach statistical significance rather than an absolute value, this is particularly so when correlation coefficient as low as  $r=.011$  which statistically would be interpreted as a poor correlation, despite a  $p$  value of .05, are quoted as demonstrating a particular tool as valid(58). Cook et al put forward the argument that validity depends on context and the construct under study and the interpretation of the results in relationship to the hypothesis put forward and must be proved each time the tool is used with a different construct(40). Downing goes further and purports that alone assessments cannot be valid nor invalid and must be interpreted in conjunction with scientific method, hypothesis and theory and the consequences to which an assessment may be put(56).

Figure 1 Schematic diagram illustrating how the differing definitions of validity are related to construct validity

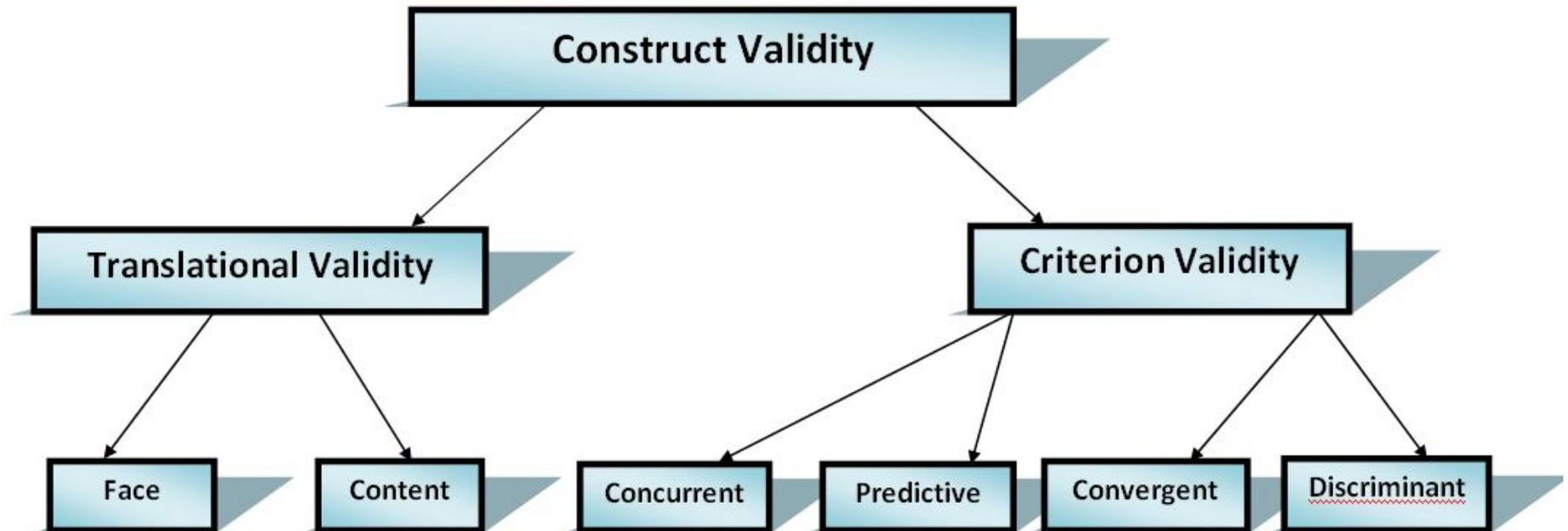


Figure 1 is adapted from DeVon et al "A psychometric toolbox for testing for testing validity and reliability"(48) Construct validity is deemed to be an overarching concept of which other forms of validity are considered lesser types of validity

## 1.2.2 Reliability

Reliability is the property of a tool or assessment to be reproducible under defined and stable conditions. It is often argued that it is an essential component to ensuring a particular tool or instrument is valid under the conditions in which it is going to be used(40, 48). Reliability is very important in assessments where observers are involved such as clinical assessments and performance based measures. Here reliability in the form of inter-rater agreement should be high in order to contribute to the validity of the instrument(45). This should be contrasted with an assessment such as MCQ's for example where a high degree of content evidence should be sought.

Performance based measures where raters are used to assess trainees using assessment tools such as PBA can be assessed using 3 main types of reliability intra-rater reliability, inter-rater reliability and internal consistency.

- Intra-rater reliability is defined as how consistently a rater can score a particular outcome or performance(45). Typically this can be assessed by the same rater watching a single procedure and scoring it on two different occasions. If the performance is subject to a practice effect such as an operation or procedure then the time span between assessments should be short in order to minimise improvement in the trainee performance but long enough to ensure the rater does not reproduce the score from memory. Waltz et al (cited in DeVon(59)) suggest an appropriate time interval between testing should be between two weeks to one month. However if the raters assess the same procedure twice (such as a single video of a performance of the same procedure by the same

person) this circumvents a practice effect which may be seen if a single trainee is performing the same procedure on multiple occasions and can reduce this as a source of error.

- Inter-rater reliability is defined as the measure of agreement between two separate raters assessing the same procedure at the same time using the same instrument. This is important when using any tool which is designed to assess performance in a psychomotor skill. Raters can introduce sources of error which include previous knowledge of the trainee being assessed or perceived difficulty of the procedure(45). A good assessment tool should compensate for assessor variance or minimise variance where possible so that it can be used to consistently assess performance across a range of trainees performing a range of procedures. In assessment tools such as PBA there can be many sources of variance; it requires an operation, which may vary in difficulty, a trainee whose performance may be inconsistent, due to being observed and a rater whose perception of the trainee and procedure may vary. The requirement of the rater to score the trainee on behaviourally defined items also can produce variance and rater training with the assessment tool can partially reduce but not completely remove this as a source of variance and hence error.
- Finally internal consistency is another form of reliability that can be measured and is applicable to assessment tools such as PBA. It is defined as agreement or correlation within instrument between items that are theoretically or conceptually related. Frequently this type of reliability is assessed using Cronbach's alpha coefficient especially in regards to assessing the reliability of questionnaires(48). However correlation coefficients have also been described and reported in the literature(58).

Statistical measurement of reliability is commonly calculated in two ways. Classical test theory (CTT) as previously discussed is a relatively straight forward process where raw data is converted to a continuous variable and comparison is made between scores by different raters to calculate a Cronbach's alpha co-efficient(43, 48). This typically gives a value between 0 and 1, where 0 equates to no concordance and 1 equals perfect agreement. An alpha co-efficient of 0.7 would indicate that reliability of an instrument or test equals 70% and 30% of the score can be attributable to error, however this value cannot tell you what the source of error may lie.

A more powerful reliability test can be performed by using Generalizable theory (GT)(60). This test requires a sufficiently large data set with multiple raters and is particularly applicable when raters are required to assess performance based examinations, such as OSCE's where there are multiple stations or multiple items within an assessment instrument. It requires specialist software and also generates a reliability co-efficient between 0 and 1. However unlike Cronbach alpha it can also highlight the sources of variance and also describe how many raters or how frequently the assessment would need to be performed to achieve sufficiently reliable results.

Both reliability co-efficient suggest values greater than 0.8 are required to reach acceptable reliability for high stakes examination, values less than this are considered only suitable for low stakes examination(43, 48).

### **1.2.3 Responsiveness**

The ability of an assessment tool or instrument to measure change over time is known as responsiveness and this is particularly important when a tool such as PBA is being used to evaluate a trainee's progress over time(46). It is a measure of how sensitive a tool may be in detecting change over time(45). Fundamental to the concept under study is the implication

that is has been designed to evaluate performance, which should theoretically improve over time and with repeated practice. If a tool cannot demonstrate responsiveness it may indicate that it is insufficiently sensitive to monitor change or that the construct under study is too simplistic and trainees are demonstrating a ceiling effect of the construct within the study(45).

Two types of responsiveness are described in the literature; internal responsiveness and external responsiveness(46). Internal responsiveness is the ability of a tool or measure to detect change over a pre specified time frame. It requires that a measure be evaluated before and after an intervention within a particular group or study cohort and is relatively easy to measure. External responsiveness is the ability of a measure or tool/ instrument to change over time in response to an external measure or standard. This example is more typically found in clinical trials. For example if patients were given a new medication to treat peripheral vascular disease and we used an new instrument to evaluate walking distance before and after the new medication was given we could evaluate that instrument in comparison to a pre-validated tool that measured health status, such as the SF36 which measures quality of life. If the changes in walking distance as evaluated by one particular instrument could show a relationship to the gold standard measurement, the SF36 in this instance, then we could say the new tool to measure walking distance is externally responsive to the agreed measure(46).

Externals responsiveness is useful when surrogate measure or outcomes are required in clinical trials, where a change in the clinical status of a patient would be evident in the measure being evaluated and the gold standard. It is of course dependant on the external measure chosen for comparison, but does mean that the measure being evaluated can then be generalised to the wider context rather than in the case of internal

responsiveness where the results are considered to be cohort specific(45, 46).

Measures of internal responsiveness can be determined by the paired t-test in normally distributed data set or the equivalent non parametric counterpart the Wilcoxon Signed Rank Sum Test. These tests evaluate the differences between the means of the cohort or study group before and after the intervention(46).

External responsiveness can be measured by relating changes in one instrument to corresponding changes in another instrument. One of the easiest ways to measure this is Pearson's product moment correlation coefficient between the measurement under study and the gold standard. High correlations i.e. values approaching 1 can be thought to indicate that the new measure is capturing change along with accepted gold standard.

However key to measuring responsiveness is that the instrument is both valid and reliable as changes in measurement can of course occur without real change in the process/ performance being measured and so demonstrating that a tool is sufficiently valid and reliable prior to assessing responsiveness is an important part of the global assessment of any instrument(46).

## **1.3 Assessment methods in surgical technical skills and simulation**

### **1.3.1 Background**

Psychomotor skills assessment of surgical trainees, has, until recent decades been a poorly examined part of surgical curricula(6). Until the advent of the New Deal in 1991 and following on from that in the form of the European Working Time Regulation initially in 2004 (EWTR) (16, 61),

the main mode of operative performance assessment was examination of log book numbers which could demonstrate the range and number of procedures in which a trainee had been exposed to but could not determine whether that trainee had acquired competence or proficiency in a procedure(62). Reliance of supervisor reports in each clinical rotation was used as a surrogate to determine if a trainee was competent to progress. Such an assessment process lacks validity and reliability and is subject to a halo effect when evaluating individual trainees(6). Competence was assumed through long hours spent on the wards and in the operating room observing and practicing operative skills on live patients with intense and frequent on calls(3, 63).

Restrictions that now exist on the amount of time spent in the clinical environment mean that time spent in the operating room is at a premium and a significant amount of training needs to be delivered though service as highlighted in the COPMED review(17) and Temple report(16). The net result of the New Deal and EWTR has meant in many instances a loss of a defined firm structure that would assume responsibility for a named trainee and support them in a structured gradual assumption of responsibility and experience(3). This no longer holds true for many junior trainees; indeed the recognition of this fact prompted the publication of the well known paper by Sir Liam Donaldson, Unfinished Business(64). This document highlighted the required expansion of the work force associated with working time restriction reforms meant that many trainees were left without adequate supervision, had poorly defined training structures that would allow them to progress in their chosen speciality and had increasing workloads that were driven by the requirements for service provision.

There have been multiple changes and reform to the training grades since this time that, some would argue, has not resulted in demonstrable improvements in the quality of training(65). However the introduction of

defined outcome led curricula which began with structured training programme such as the Orthopaedic Competence Assessment Project (OCAP) in orthopaedic surgery, which has now been amalgamated with the Intercollegiate Surgical Curriculum Project (ISCP) for general and speciality trainees, now means there is at a nationally agreed system and process to train and evaluate performance.

A number of external factors have altered how we train the present generations of surgeons. The lack of access to operative lists due to service commitments and the requirements of service pressure to complete operating lists on time<sup>(65)</sup> have reduced operative exposure for the individual trainee. Coupled to this, associated advances in medical science that mean that many disease processes are treated medically rather than surgically. Increasing numbers of trainees, which are now required to ensure rotas are EWTR are compliant, mean there are fewer operating lists available for each trainee to attend, eroding the acquisition of surgical skills. These changes have introduced the ethos of making every training opportunity count and of finding new ways to train the next generation of surgical trainees such as simulation and deliberate practice outside the clinical environment (16, 20).

Assessment of any form is a labour intensive process, especially in evaluation of surgical performance, where assessors need appropriate insight in to what constitutes an empirically and technically correct performance<sup>(63, 66)</sup>, unlike for example a written examination which can be delivered en masse to many trainees all at once. A number of assessment methods are in clinical practice at present. Some of these methods have little value outside of the research environment or simulation suite and their use is solely restricted to evaluating performance in a rater free manner. They are designed to provide assessment and feedback on performance without the presence of a trainer and include

processes such as simulated metrics and computed analysed efficiency of motion(67). Others require the presence of an assessor, preferably with suitable experience in the procedure being assessed and generally consist of checklists and rating scales of varying length and structure(68). A systematic review from 2011 identified 106 unique articles which evaluated multiple assessment tools which required the presence of a human assessor and found 29 studies which used global assessment tools, 30 used procedure specific or operations specific tools and 47 studies that included a combination of global and task specific or operation specific methods(69). Given the wide scope and heterogeneity of the assessment tools and studies only those that have historical context or are common use are described in more detail in the following subsections.

### **1.3.2 Rater free assessment methods**

Typically assessment methods that do not require a human assessor are achieved through computer generated data, which can compare or measure performance in relationship to a pre-programmed ideal standard. Simulator metrics refers to a digital read out of performance as defined by pre-programmed software contained within high fidelity simulators. The more advanced simulators can produce structured training programmes and monitor progress which can be read immediately or downloaded for future reference. Simulators of this kind include laparoscopic, endoscopic, and endovascular high fidelity simulators such as the ANGIOMentor. A typical read out from an endovascular simulator is illustrated in figure 2 (p 38). They often include time for completion of procedure, type or number of errors and in the case of endovascular simulators such as the ANGIOMentor give the amount of contrast use or fluoroscopy use. They may also give feedback on the quality of the final product such as percentage of lesion covered or residual stenosis in the case of angioplasty and stenting procedures. Laparoscopic virtual reality trainers include such

devices as The Minimally Invasive Surgical Trainer Virtual Reality (MIST VR)(6). This is a computer generated 3 dimension environment, which trainees interact with a simulated procedure using real laparoscopic instruments which have motion sensors at the tip. Data is then gathered by a computer and a score is generated that monitors performance. In isolation this may not be useful for the trainee without specific feedback on performance errors but a number of studies(70) (71)have show this to be a valid tool which produces reliable and reproducible data. While such assessment methods may be an attractive option for assessing trainees there are limitations to simulator metrics use. The benefits of such simulator metrics or efficiency of motion scores is that they can be standardised and reproduced but are more suited to experimental studies or simulators where the environment and procedure being studied is strictly controlled as opposed to the operating theatre where the same operation will vary from patient to patient. Additionally simulators may over estimate or pick up clinically irrelevant errors and it is difficult for an inexperienced operator to understand which are the errors that are important and those that can be safely ignored. For this reason trainees still need a level of experienced supervision to provide guidance and clinically relevant feedback(39).

Figure 2 Readout from the ANGIOMentor endovascular simulator for superficial femoral artery angioplasty

ANGIO Mentor		Simbionix	
<b>Performance Sheet - Instructor Instructor</b>			
Module: SFA Intervention Case: 1			
Date: 5/17/2012 1:54:37 PM			
Session Date		05-17-12	
Total procedure time		13:46:25	
Procedure Log		00:06:28	
Total amount of contrast injected (ml.)		INFO	
Total fluoroscopy time		29	
Total amount of time that patient was exposed to fluoroscopy.		00:03:40	
Number of TSA sequences recorded		3	
Interventional procedure(s)		1	
The number of interventional procedures (Predilation/Stenting/Postdilation) that were performed during the case.			
Number of guide wires used		4	
Total number of guide wires that were introduced into the patient's body during the procedure.			
Number of Dx catheters used		4	
Total number of diagnostic catheters that were introduced into the patient's body during the procedure.			
Residual stenosis (%)		23	
The percentage of stenosis that remained at the end of the case.			
Residual transstenotic pressure gradient (mmHg)		15.07	
The difference between the blood pressure measured on both sides of the stenosis, at the end of the case.			
Time to obtain Dx aortogram		00:01:01	
Time elapsed between the beginning of the procedure to the time the diagnostic image of the aortic bifurcation has been acquired.			
Time to position guiding catheter or sheath ready for intervention		00:04:34	
Time elapsed between the beginning of the procedure to the time the first balloon or stent is introduced.			
The duration of time the predilation balloon was opened		00:00:54	
The inflation time of the predilation balloon within a lesion.			
Correct predilation balloon diameter		Yes	
Yes = when the diameter of the balloon is at least 1mm smaller than the native vessel diameter. Applies when the choice of the balloon diameter is intended to enable stent passage through the lesion.			
Correct PTA balloon diameter		No	
Yes = when the diameter of the balloon is equal to the native vessel diameter (it may be up to 0.5mm smaller or larger). Applies when the choice of the balloon diameter is intended to resolve the stenosis.			
Percentage of lesion length covered by the predilation balloon(s)		100	
Cumulative coverage percentage of predilation balloon(s) inflated within a lesion. The referred lesion length is the lesion's section which its proximal and distal diameters are the average between the most stenotic part of the lesion and the native vessel diameters.			
Difference between predilation balloon length and lesion length (mm)		20.10	
The difference between the length of the predilation balloon and the length of the lesion in which it was inflated. The referred lesion length is the lesion's section which its proximal and distal diameters are the average between the most stenotic part of the lesion and the native vessel diameters.			
Difference between predilation balloon diameter and vessel diameter at lesion site (mm)		-1.01	
The difference between the diameter of the predilation balloon and the average native vessel diameter. The referred lesion diameter is the average between the proximal and distal native vessel diameters.			
Advancing guide or sheath without a leading wire (mm of travel)		58.15	
Total distance in millimeters of advancing guide(s) or sheath(s) without a leading wire.			
Patient experienced hemodynamic irregularities during the procedure (YES or NO)		No	
Patients experience symptoms when their heart rate or systolic blood pressure fall below threshold values.			
Lowest sustained HR during procedure (beats/min)		88	
Lowest recorded HR during procedure that was sustained for more than 60 seconds			
Patient HR at end of case (beats/min)		88	
Highest sustained BP during procedure (mmHg)		182	
Highest recorded systolic BP during procedure that was sustained for more than 60 seconds.			

A number of additional rater free assessment methods are described in the literature(5, 6, 67). Efficiency of motion analysis developed in the Imperial research labs in London has been designed to help facilitate research into laparoscopic, open and in some instance live operation skills assessment(72-74). The unit is bespoke and consists of motion tracking transmitters which a subject may wear on the back of their 2<sup>nd</sup> metatarsal, often secured with a surgical glove to prevent slippage. The camera's within the unit can then monitor the transmitters and calculate x, y and z co-ordinates which is accurate to within 1mm and calculate the number of hand movements to give a surgical efficiency score for each procedure being monitored. Like simulator metrics such assessment methods cannot distinguish between what may be relevant or non relevant errors. Often such assessments are combined with a final product analysis which still requires the presence of a rater.

Datta et al were able to validate this dexterity analysis tool in bench tasks performing bowel anastomosis and vein patch repair of an artery and predictably experienced surgeons outperformed junior trainees(72) however like the MIST-VR assessment technique speed and dexterity are not necessarily measures of technically correct performance. Other studies exist which confirm validity for time and number of movements with senior operators outperforming junior trainees(68). But it has only been used once in operating rooms where the assessment method was found to have concurrent validity( $r = <0.63$ ) in correlation with the assessment tool known as the Objective Assessment of Technical Skills (OSATS) and only moderate inter-test reliability ( $r = 0.63$ )(75).

The Advanced Dundee Endoscopic Psychomotor Trainer (ADEPT) is an infrared system that utilises infrared sensors attached to the surgeon's hands and using surrounding infrared cameras the position of the trainee or surgeon's hands can then be extrapolated to monitor movement. Like

the previously mentioned systems it is a form of motion analysis but it is limited by signal disturbance from line of vision in the sensors or signal overlap if sensors are placed on both hands. It has not been well validated and there are few studies documenting its use(6).

There are a number of other motion tracking systems which include the ProMIS™ , HUESAD and TrENDO. The majority of these studies are limited to validation of the device itself using time as the discriminator and there is little validity data on their usefulness in assessing or discriminating between trainees(68). All these rater free assessment methods are summarised in table 1 (page 41).

**Table 1 Summary of the main assessment devices which provide rater free feedback during assessment of surgical psychomotor skill**

<b>Assessment Device</b>	<b>Assessment Method</b>	<b>Evidence for Validity and Reliability</b>
<b>Minimally invasive surgical trainer- Virtual Reality (MIST-VR)(25, 27, 28)</b>	Total 3 dimensional immersion simulation using a laparoscopic simulator with motion sensors at the tip to monitor performance.	Shown to be valid and reproducible.
<b>Efficiency of Motion (Imperial Research labs) (ICSAD)(72, 74)</b>	Motion tracking sensor on the back of each 2 <sup>nd</sup> metatarsal to calculate x,y,z coordinates to track hand movements to calculate a surgical efficiency score	Demonstrated to be valid for time and number of movements between junior and senior operators. Moderate inter-test reliability in the operating room.
<b>The advanced Dundee Endoscopic Psychomotor Trainer (ADEPT)(76-78)</b>	Utilises and infra-red system of cameras and sensors to monitor movement.	Not well validated only 2 studies confirming its use, which discriminate between master surgeons and junior trainees
<b>ProMIS(TM)(79)</b>	Hybrid, virtual reality laparoscopic trainer (Not mechanism for assessment)	Demonstrated to be partially valid, and have good inter-rater reliability for a mixture of tasks
<b>Hiroshima University surgical assessment device (HUESAD)(80-83)</b>	Electromagnetic motion tracking device, using laparoscopic instruments and calculated path length.	Shown to be construct valid between novice and experts
<b>Tracking Endoscopy (TrEndo)(84, 85)</b>	Utilise the fulcrum of the laparoscopic instrument to track motion, limited to minimally invasive surgery	Demonstrated construct validity between the expert and novices for square Knot tying

### **1.3.3 Assessment Methods Requiring a Rater**

Assessment methods that require a rater or assessor are more commonly used in clinical and day to day practice. A number of different assessment tools have been described that have been designed to be used to assess surgical performance. The majority are experimental tools but a small number are in regular clinical use, including the assessment method known as PBA, which is the focus of this overarching study. The studies are summarised in table 2 (p43) and discussed in further depth in the following subsections, where they are further divided into assessment methods with historic context, experimental tools and finally those assessment methods that are in routine clinical practice.

**Table 2 summary of surgical psychomotor assessments which require human rater input**

Assessment Method	Skills and Mechanism of Assessment	Validation and Reliability
<b>Kopta's Observation guide</b> (Historical)(5)	80 criterion based on explicit behaviours marked on a likert scale of 1 to 3	Not validated or assessed for reliability
<b>Global Operative Assessment of Laparoscopic skill (GOALS) (86-88)</b> (Experimental use only)	Very similar to the OSATS tool with a 10 item checklist and global skills applicable to laparoscopic surgery, such as depth perception and bimanual dexterity	Found to be valid but reliability not assessed. Used to demonstrate progression in the "Fundamentals of laparoscopic surgery" and skills transfer
<b>Observational Clinical Human Reliability Assessment (OCHRA)</b> (Experimental use only)(89)	An iterative assessment method developed using video analysis of Laparoscopic cholecystectomy to identify common clinical errors	No validity or reliability analysis performed.
<b>Imperial College Evaluation of Procedural Technical Skills (ICEPS)</b> (Experimental use only)(90-92)	Developed specifically for simulated sapheno-femoral junction ligation, for seven skill domains, marked on a Likert scale of 1-5	High reliability, but validity limited to junior surgeons in simulated and live procedures.
<b>Direct Observation of Procedural Skills (DOPS) (93-95)</b> (Used in clinical practice)	Short checklist of explicit behaviours, marked as satisfactory or development required. Followed by a global summary score anchored by explicit descriptors	Not well validated or found to be reliable.
<b>Objective Structured Assessments of Technical Skills (OSATS) (68)</b> (Used in clinical Practice)	Two part assessment, checklist of key skills to the index procedure assessed and global rating scale for key operative skills marked on a Likert Scale from 1-5	Multiple studies assessing construct validity in simulation, bench top procedure and live theatre based assessments
<b>Procedure Based Assessment (PBA) (58, 96, 97)</b> (Used in clinical practice)	Two part assessment method, extensive checklist covering, consent, pre-operative planning, intra-operative steps, access and closure, and post procedure management, marked either as satisfactory/ or development required. With a simple five part global summary score anchored by explicit descriptors.	Minimal studies assessing validity and reliability in simulation. One large prospective study in the workplace assessing validity and reliability in live operative performance.

### **Kopta's Observation Guide**

One of the earliest forms of psychomotor skills assessment is described by Kopta in 1971(67). He observed 40 years ago that surgical technical skills were poorly evaluated in any kind a systematic manner. He published a paper describing a guide which was divided into 5 domains looking at team working, cognition, checklists of procedural skill, outcome and an evaluation of any critical incident through the procedure. In all, 80 criteria were defined with each being marked on a Likert scale of 1-3 by a rater. There is no record of whether this guide was validated further(5).

### **Experimental surgical assessments tools**

In addition to Kopta's Observation guide above there are a number of other assessments tools that have been published in the literature on surgical assessment. The first of these GOALS or Global operative assessment of Laparoscopic skill was designed by Vassiliou et in 2005(86) which described a tool not too dissimilar to the OSATS which was used to assess performance during laparoscopic cholecystectomy and laparoscopic appendicetomy in 21 trainees in a paper published in 2007 by Gumbs et al(87). Like the OSATS it has a global rating scale from 1-5 in the domains of tissue handling, depth perception, bimanual dexterity, efficiency, autonomy and level of difficulty and a 10 item checklist. They were able to demonstrate construct validity for surgical performance for the entire procedure but did not evaluate reliability.

Tang and colleague developed a method of clinical error assessment based on video analysis of 60 surgical trainees performing laparoscopic cholecystectomy called Observational Clinical Human Reliability Assessment (OCHRA)(89). Using these video's they were able to identify technical common errors and then introduce corrective training if required. The limitations of such an assessment method include its retrospective nature and labour intensive analysis. There is a lack of data on validity and

reliability but such a mechanism of assessment is very useful for formative feedback to the trainee and can help design curriculum for specific training programmes.

The Imperial College Evaluation of Procedural Technical Skills (ICEPS) is a procedure specific assessment scale which has been described for use in simulated saphenofemoral junction ligation(90, 92). It is designed along the same lines as the OSATS with each domain being rated on a Likert scale of 1-5, but contains 7 domains that are specific to the procedure being assessed. It is not well described in the literature despite it possessing high reliability (Cronbachs alpha 0.96). However in a single study using video blinding methods of 50 surgeons comparing ICEPS with the OSATS tool it was not demonstrated to be construct valid, except between SHO's and junior registrars. The alternate subgroups within this study, which included senior registrars, new consultants and senior consultants, demonstrated no significant differences in scores although there was a gradual trend for increasing score with seniority(90). This result however may be skewed once again by the ceiling effect. Saphenofemoral junction ligation is a relatively simple procedure and expecting significant differences between senior registrars and consultant may well be ambitious. However a further study which utilised the ICEPS tool in assessment of skills transfer between a simulated saphenofemoral junction ligation in SHO's to live performance it was demonstrated clear and significant differences for all seven domains, suggesting certainly at a junior level of training this could be a useful and reliable instrument(92). (Appendix 1)

### **Directly observation of procedural skill (DOPS)**

The original DOPS assessment method preceded the introduction of the ISCP website in 2007. Initially paper forms which were completed by educational supervisors on minor procedures such as arterial punctures, cannulation and urinary catheterisation skills performed by trainees, they started to become routine after the introduction of foundation training.

Within the ISCP web environment they have now been modified and do not look dissimilar to short procedure based assessment forms and are designed to reflect assessment or part of a complex procedure such as a carotid endarterectomy or less complex procedures such as central venous cannulation. (Appendix 2)

The evidence for validity and reliability of DOPS is scanty. Wilkinson in 2008(94) concluded that there was no validated procedural performance assessment in the literature and subsequent reviews in simulated procedures in medical students suggest reliability was low and they only possessed face and content validity(95) and no evidence to suggest that they improve performance(15). Naeem in 2013(93) concluded through narrative review that there was no evidence that it was a useful workplace based assessment.(98)

### **Objective Assessment of Technical Skills (OSATS)**

First designed in Canada, the Objective Structured Assessment of Technical Skills (OSATS) is now well established as a gold standard method of assessment surgical skills. The first paper describing its use was published in 1996 and outlined the original form which consisted of a procedure specific checklist and global rating scale with 7 domains of surgical skill; such as flow of the operation, efficiency of motion, tissue handling skills and knowledge of the operations (appendix 3 )(5, 99, 100). Each domain is marked on a Likert Scale of 1-5 anchored by explicit descriptors. This initial study described a simple correlation study between OSATS scores and faculty rankings for 6 junior and 6 senior surgical residents and found high correlation between OSATS scores and senior resident rankings, but not for junior residents(99). The same group then designed a study to assess surgical interns performing a variety of surgical procedures on bench models which included synthetic and live porcine tissue. Twenty residents performed across a 6 stations including excision of skin lesion, hand sewn bowel anastomosis, inferior vena cava haemorrhage, abdominal wall

closure and insertion of a T-tube. This study assessed the construct validity using year of training as the variable and inter rater reliability and inter station reliability. Construct validity was found for both the live and bench formats with respect to year of training using the global rating scale, but only a marginal effect on checklist rating scores in live models and a non significant effect with checklists in the bench format. Inter rater reliability across six stations as measured by Cronbach's alpha was 0.74 for the bench model and 0.66 for the live format using global rating scales. For across the same six stations the checklist element reliability was estimated at 0.33 in the bench models and 0.61 in the live format. Inter rater reliability between assessors for the global and checklist ratings was medium to high and ranged from 0.64 to 0.72 across both the bench and live model formats. The main finding from this study was that bench models gave very similar results to the live format and could be used to evaluate residents using this cheaper and reproducible option and that this assessment method was valid and reliable, certainly in the global rating section, in evaluating simulated psychomotor performance(101). Reznick et al went on to assess a larger cohort of trainees on the same bench models and found that the tool was construct valid for training level and had a inter-station reliability of 0.78 for checklists and 0.85 for the global rating scale providing evidence that it could be a valid a valid and reliable assessment for simulated bench models(102).

The OSATS assessment has been used in multiple studies of surgical skill and simulation validation procedures since it was first described and published(68). Those studies of historical context and relevance to vascular surgery are summarised in table 3 (page 50-51). The original format and modified versions have been used to assess performance in open(58, 102-107), laparoscopic(103, 108, 109) and endovascular simulations studies(110-112). Van Hove in 2010 published a systematic review of assessment methodologies covering surgery and gynaecology skills

assessment and found 26 studies using OSATS as the assessment method. Nineteen studies were based in a skills laboratory setting. The remainder were in an operating theatre environment. In those studies where inter-rater reliability was established both in the lab and theatre setting it is over 0.80 in 10 of the studies included in the review where the majority were used to assess performance in gynaecology bench tasks and operations. However of these studies there were only three that focused on vascular skills; 2 which focused on video assessment of a skills lab based assessment of a vascular anastomosis and patch(107, 113) and the remaining study was performed in the operating room during live carotid endarterectomy procedures(106). Only one of these studies provided inter-rater reliability data which was over 0.8 (Cronbachs' alpha) for the checklist and global rating scale element for each form(107). Since the publication of this review two further studies have been published which utilise simulated vascular procedures. In 2011 Price et al described a study where 39 trainees were randomised to either expert tutorial versus expert tutorial and self directed practice on a micro vascular anastomosis, with evaluation at baseline and after the intervention. The OSATS tool proved to have an intra-rater reliability of 0.8 with blinded assessors and was demonstrated to be internally responsive. In a un-blinded study Duran et al in 2013 went on to assess 92 trainees, representing one of the largest studies to date with the OSATS tool, performing an end-to side vascular anastomosis. They found it had a high intra-rater reliability of the global element and high internal validity between the global and checklist elements of the tool. They did not assess the reliability of the checklist as they found it difficult and cumbersome to use and it had high internal validity with the global element of the checklist. The global element was also construct valid between novice, intermediate and experienced trainees. Few of these studies are considered to be high level evidence with minimal evidence of blinding or randomisation but where the evidence is described it suggests

that the OSATS tool can be a reliable and valid measure of performance in vascular procedures, certainly with respect to training year and potentially prior experience as well. It also suggests that the global rating scale of the OSATS tool is potentially the most reliable, valid element and acceptable part of the assessment.

The global element of the OSATS tool has also been studied alongside the PBA tool in the live theatre environment, across a range of procedures, in a large prospective study. This study also demonstrated that the global element of this tool was potentially more reliable (reliability co-efficient 0.83 vs 0.81) than the checklist element of the tool, when used analysing single index procedures(58), but reliability dropped significantly when it was used across a range of procedures.

While the evidence base in open vascular simulation is small a considerable number of studies exist, especially in live and simulated gynaecology procedures, and so it has been adopted as the obstetrics and gynaecology work place based assessment method of choice for operative performance in the live theatre environment in the United Kingdom. Examples of such forms can be found on the Royal College of Gynaecology and Obstetrics website(114).

Unlike such assessment methods as computer generated metrics and efficiency of motion analysis the presence of a live rater is essential to its use and each and every evaluation. So while it may be labour intensive it does provide the trainee with immediate and hopefully relevant feedback which is readily understandable and provides specific areas for improvement(63).

**Table 3 Summary of the main historical studies and relevant vascular surgical studies utilising Objective Structured Assessment of Technical Skills**

Study	Setting	Outcomes	Evidence for Validity and Reliability
<b>Faulkner 1996(99)</b>	12 (6senior and 6 junior) residents performing 6 simulated bench and live general surgical tasks	To assess concurrent validity between faculty rankings and OSATS scores.  High correlations found for global score in senior resident (0.66-0.71) , but not for junior residents (-0.2)	Checklist elements demonstrated t not to be valid in junior residents in live and bench formats. Only the global element was valid in senior residents. Some raters declining to rank junior residents, who had minimal experience at these skills.
<b>Martin 1997(101)</b>	22 general surgical residents assessed in a bench format or live format of general surgical skill	Demonstrated equivalence between bench and live scores. Intra-rater reliability for OSATS for checklist and global scores was moderate to high (0.61- 0.74) except the checklist for the bench format.	Validity only proved for global rating scores in bench and live formats. The checklist element was only significant between years 1 and 5 for the live format and marginal for the bench models.
<b>Reznick 1997(102)</b>	48 residents assessed using checklist element and global rating scores of the OSATS performing bench model simulation of general surgical tasks	Inter station reliability and large scale construct validity of bench station examination.	Inter-station reliability for checklists 0.78 and global rating 0.85. demonstrated to be construct valid for training level by ANOVA for global and checklist scores
<b>Datta 2002(107)</b>	Vein patch anastomosis in 50 subjects from basic surgical trainees to consultants	To assess concurrent validity with motion tracking and the OSATS tool and construct validity of the OSATS tool	Concurrent validity (spearman $r = 0.53$ $p < 0.01$ ) between OSATS global rating and motion analysis. Construct valid for each experience group for global ratings only (Kruskal-Wallis $P = < 0.001$ ) not for checklists.
<b>Brydges 2007 (113)</b>	18 junior and 9 senior residents performed a live porcine model end to side vascular anastomosis, before and after study assessed using the ISCAD and OSATS then assessed on the final	To demonstrate convergent validity with ISCAD and OSATS between junior and senior residents, and construct validity. Blinded assessment using video	Construct validity for the OSATS was demonstrated for senior and junior trainees ( $p = 0.49$ for global rating and $0.07$ for checklists) (Mann-Whitney U).  Also convergent validity with ISCAD

**Table 3 Continued**

	performance		
<b>Beard 2007(106)</b>	17 trainees and 11 consultants performed carotid endarterectomy in live theatre setting over 2 years	Assessed construct validity between trainees and consultants and for previous experience for the checklist element and global rating scales.  P=values were significant for task specific p=0.031 and global rating p=0.001 between trainees and consultants.  Validity for previous experience in trainees was significant for task specific (r= 0.83) and global scores (r=0.69). but only global scores were significant ("p" value <0.05) for consultants (r=0.82)	Single observer  Demonstrated construct validity between trainees and consultants for training level for global and checklist scores.  Only construct valid for prior operative experience in trainees.
<b>Duran 2014(115)</b>	Ninety two trainees assessed performing an end to side anastomosis using a modified OSATS tool with checklist and global rating scale(GRS)	Internal validity between checklist and GRS was 0.71. Between items in the GRS was 0.92. Reliability between raters was 0.64-0.77 for each item within the GRS.  Construct validity using one way analysis of variance between novice, intermediate and experienced trainees as determined by training level p=0.001	Assessors found the checklist cumbersome and difficult to use, therefore did not assess reliability of the checklist.  Un-blinded  Experience in the procedure inferred by training level.
<b>Price 2011(116)</b>	Microvascular anastomosis on a procedural trainer. Randomized to expert tutorial and additional self directed practice in 39, first and second year surgical trainees, assessed again on live porcine model	Assessed using a traditional OSATS tool.  Only reliability assessed between assessors, intra-rater reliability 0.8 (intra-class correlation co-efficient)	Randomised and blinded assessors

## **Procedure Based Assessment (PBA)**

PBA became the official method of comprehensive skills assessment for the surgical trainee in the United Kingdom in 2007. This is when the use of the ISCP became compulsory for all surgical trainees in a recognised surgical training post(117). The new curricula at this time needed relevant assessment elements to ensure the curriculum outcomes were being met and PBA was designed to map to the curriculum objectives and to assess the psychomotor skill elements required. (Appendix 4) It is a form of varying length depending on the complexity of the procedure which consists of 2 elements. The first part is a checklist of competencies that are marked as "not seen" (N), "development required" (D) and "satisfactory for completion of CCT" (S). The checklists of competencies are divided into a number of domains: pre-operative planning, consent, pre-operative preparation, exposure and closure, intra-operative technique (generic and procedure specific) and post operative management. The second part of the form includes a global summary or level score which allocates the trainee an overall mark from 0 (insufficient evidence to perform the procedure) to level 4 (able to complete the procedure unsupervised and could deal with any complication that arose). Unlike the OSATS before it there is minimal information in the public domain regarding its development and psychometric properties. Assessment of face and content validity had been derived by an iterative process between the royal colleges(4) but there were initially no validity and reliability studies performed on PBA as it is now or as it was then in 2007. Contemporary studies on PBA are summarised in table 4 (pp 58-61).

The PBA form and indeed many of the work placed based assessment had a less than encouraging beginning with many trainees dissatisfied by this form of assessment. There were technical issues of the internet based form, lack of supervisor training and general difficulties in getting the forms

complete in a timely manner(118, 119). Designed as an assessment for and of learning this was infrequently the case and many trainers and trainees saw such forms as a tick box exercise. Getting assessments performed still remains problematic from time to time, but the ethos towards such forms is changing as more trainees and trainers become familiar with the process.

The earliest iteration of PBA is perhaps described in the study based in the Trent deanery in 2003 whereby a similar form, not too dissimilar to the present Direct Observation of Procedural Skills (DOPS) form, was first described and evaluated(120).

This study essentially described a form with 19 core skills as a checklist for each index procedure outlined in an operative competence form. The 19 core skills are outlined as generic rather than procedure specific as can be seen with the modern PBA form. However in addition to the core skills checklist a overall rating was awarded not dissimilar to the global rating seen on the PBA but instead of trainees being rated from 0-4 they were rated from A to D (A-able to teach the procedure, B- able to perform the procedure unsupervised, C- able to perform the procedure with supervision to D – unable to perform the entire procedure). For the purposes of analysis this was converted to a numerical value. Twenty three higher surgical trainees were prospectively assessed in this way performing key index procedures. Scores were correlated and analysed using spearman's correlation coefficient but unfortunately mean rather than median scores were quoted. However what the authors did show was that mean operative competence scores correlated with duration of surgical training ( $r=0.69$   $p=0.01$ ) Similar results were shown for total number of procedures performed and mean competence scores. Of the index procedures described there were only 6 procedures that the trainees had performed in sufficient numbers to analyse independently and of these, appendicetomy, inguinal hernia repair and varicose veins showed no

correlation between numbers of procedures and mean competence scores due to a ceiling effect; the trainees had all achieved the maximum score possible for these procedures. Only the more complex procedures such as left colectomy were able to demonstrate a correlation between mean competence score and operative numbers  $r=0.73$   $p=0.01$ . Trainer bias and a potential halo effect were potential confounding factors in this study as the assessors were the trainees' nominated educational supervisors and also admitted to assessing the trainee without actually having seen them perform some of the procedures.

As PBA become further incorporated into training programmes, researchers began to use the PBA tool in educational studies. The first of these studies described in the literature, both incorporate an element of simulation. In 2007 Sarker et al published the results of a moderate sized validation study using 28 trainees performing 84 live (supervised) and 112 simulated (unsupervised) laparoscopic cholecystectomies using PBA as the assessment tool to evaluate outcome measures in a blinded manner using video assessment(96). Only the itemised checklist part of the PBA tool was used and was divided into the generic competencies and procedure specific competencies (similar to those in appendix 4) .The PBA specific tool for laparoscopic cholecystectomy was determined to have a mean inter-rater reliability of Kappa =0.86 for live and 0.84 for unsupervised operations. Mann-Whitney U analysis in determining differences between junior and senior higher surgical trainees (HST1-4 & 5-8) was significant for generic competency items for live and unsupervised operations. However construct validity for the specific competency items was only significant for the unsupervised simulated procedures and not for the live supervised operations. The authors speculated that this abnormality was due to correction of junior trainees techniques during the live operations by supervisors and because of the nature of the blinding technique (video

analysis) it was not possible to determine how much input the supervisor had in the live operations.

The second study described in the literature is a skills transfer study utilising the PBA tool from the OCAP website for knee arthroscopic(97). Here 20 junior trainees who were procedure naive to arthroscopy were randomised using sealed envelopes to training on a bench top arthroscopic simulator (three sessions of six simulated arthroscopies) or no training. Each trainee was then asked to perform an arthroscopy for the first time in theatre after traditional instruction and demonstration of the procedure. They were then assessed by a supervisor who was blinded to their training status. Both the procedure specific intra-operative elements from the relevant PBA tool and modified OSATS global rating scale were used to assess trainee performance. The PBA tool was able to distinguish between those that had received simulator training and those that had not. Levels of significance for the procedure specific element were  $p= 0.0007$  and for the global rating scale  $p= 0.0011$  demonstrating that the simulator could provide a significant learning benefit for junior trainees.

In 2010 Beard et al published the result of a two year study on the validity and reliability of PBA in the work place(58). An observational study across 3 teaching hospitals in Sheffield, it has been one of the few multicentre trials of such an assessment method. Prospectively designed, it was able to gather data from 749 PBA assessments across a range of specialities including, general, orthopaedics, gynaecology and cardiothoracic surgery by multiple experienced, trained raters. From this study the authors were able to extract data to confirm that PBA was a valid and reliable assessment method for psychomotor skills assessment in surgical trainees in the live theatre environment. Unlike many smaller studies on reliability which utilised classical test theory, which is limited by the data set in which it is gathered, they were able to construct a d-study and derived reliability

coefficients using generalizable theory which allowed them to model reliability based on the number of procedures observed or number of ratings performed. An example of this is illustrated in Table 5 (p59) This allows the data from this study to be extrapolated beyond the data set of study to a population as a whole. Despite this there are limitations to this study.

The study was conducted in a teaching hospital where the study population tended to be either very junior or senior trainees, there were few intermediate level trainees in the data set gathered. The study also used data from gynaecology trainees of which on retrospective analysis 42% had training concerns, which may result in under estimation of some for the validity and reliability data.

The number of individual trainees per speciality was small. In total there were 85 trainees with only 11 in vascular training posts and the majority from obstetrics and gynaecology, 33, which given the noted training concerns may have affected the data and underestimated the reliability. Post hoc analysis of comparison of PBA performed within the O&G speciality versus non O&G showed reliability to be even higher when those cases performed by obstetrics and gynaecology trainees were selected out. The PBA form achieved acceptable reliability  $>0.8$  using only 2 assessments of per index procedure. (Table 5 p 62)

Construct validity in this study was examined by correlation of PBA scores with number of years in UK training, number of procedures previously performed and age. Significant correlations were found between the checklist elements and level scores of the PBA form and ST level, age, and number of years in training. Most significantly the study group was able to demonstrate a significant correlation with number total number of index procedures previously performed and recent number of procedures performed with co-efficients approaching  $r= 0.5$   $p<0.001$  for the global

level score demonstrating the hypothesis that PBA shows construct validity for previous operative experience; that the PBA is measuring competence achieved rather than just reflecting time in training.

**Table 4 Summary of the relevant studies historic studies utilising Procedure Based Assessment**

Study	Setting	Outcomes	Evidence for Validity and Reliability	Study Comments
<b>Thornton et al (120)</b>	Trent deanery higher surgical trainees assessed by their educational supervisors using a pilot form very similar to the modern day PBA on a range of general surgical procedures	To assess the validity of the form to training year for each of the general surgical procedures included.	Found to be valid for trainee year and total number of procedures previously performed, but not for index specific procedures, such as appendicetomy, inguinal hernia repair and varicose veins	Less complicated procedures found to have a ceiling effect. Un blinded assessors with no direct evidence that the assessors had seen the trainees actually perform the procedures outlined in the study.
<b>Sarker et al(96)</b>	Laparoscopic Cholecystectomy, 84 live and 112 simulated procedures assessed by 28 trainees assessed using competency checklist only- separated into generic and specific technical skills	Inter-rater reliability assessed using Kappa for each of the competencies and construct validity between junior and senior residents..	Construct validity between novice and experienced operators was demonstrated for live and simulated operations for generic skill utilising Mann Whitney U, But only construct valid for specific technical skills (not generic) for simulated operations. Inter-rater reliability was 0.86 for live operations, and 0.84 for simulated.	Only the checklist part of the PBA tool was used. The Global summary score was not used.Only a single procedure assessed: laparoscopic cholecystectomy  Blinded assessment of the cases using retrospective review of video footage.

**Table 4 Continued**

<p><b>Howells et al(97)</b></p>	<p>20 procedure naive trainees randomised to receive training on an arthroscopic simulator (18 simulated knee arthroscopies) vs no training, prior to assessment during a live arthroscopic procedure by a rater blinded to training method</p>	<p>Designed to demonstrate transfer of skills from the arthroscopic training model to live patients</p>	<p>Indirect construct validity proved as prior experience demonstrates superior performance.</p> <p>Significant improvement demonstrated using the procedure specific (<math>p &lt; 0.001</math>) and global summary score (<math>p = 0.001</math>) between the two groups.</p>	<p>One of the few studies which demonstrate transferability of simulator skills to the theatre environment.</p> <p>Blinded assessment and Limited to a single procedure, with only 10 trainees in each arm.</p>
<p><b>Beard et al(58)</b></p>	<p>Based in 3 south Yorkshire training hospitals. 749 PBA assessments were performed to evaluate trainees across a range of specialities, general surgery, orthopaedics, vascular and obstetrics and gynaecology and cardiothoracic. Live operations were assessed by independent trained assessors and the trainee supervisor</p>	<p>To assess validity and reliability of PBA in the theatre in environment across a range of specialities and training grades.</p> <p>Utilising to main surgical psychometric assessment tools the PBA and OSATS tools.</p>	<p>Compared performance with OSATS, but did not assess concurrent validity</p> <p>Found to be valid for age, training year, previous operative exposure for both the checklist and global summary elements across a range of specialties, as demonstrated by Pearson correlation co-efficients</p> <p>Found to be highly reliable for inter- rater reliability utilising a d-study</p>	<p>Prospective, well funded large study with robust methodology</p> <p>Generated a data set which could be used to shape assessment protocol in a wider context. But again the PBA level or global summary score was found to be the most reliable.</p>
<p><b>Osborne et al (121)</b></p>	<p>25 trainees in the first 4 years of speciality training, used</p>	<p>Main aims were to assess construct validity and to compare self and</p>	<p>Construct validity for year of training for self assessment was</p>	<p>Single centre, small study, which successfully</p>

**Table 4 Continued**

	<p>PBA to perform self assessment for appendicetomy, which was then compared to an external assessment. Quantitative and Qualitative analysis of self assessment and external consultant assessment</p>	<p>external rating of PBA.</p> <p>Also qualitative analysis of performance comparing trainee comments with assessor comments</p>	<p>demonstrated.</p> <p>No significant differences between self and external global summary scores, although a small proportion of trainees over rated themselves on the six checklist sections, there were no under-ratings.</p> <p>No significant difference between learning needs identified between self and external assessment. Self assessment was more likely to identify non-technical skills, situation awareness, decision making and leadership. Vs communication skills from external assessors and technical skills</p>	<p>identified that trainees have insight in to their technical skills which closely matches that of the external consultant assessment.</p> <p>Trainees also easily identify technical training needs that correlate well with external assessors.</p>
<p><b>Duschek et al(122)</b></p>	<p>3 day intensive training course, which assessed 10 trainees before and after simulated practice at carotid patch plasty. PBA and technical quality were used as</p>	<p>A before and after study where PBA was used to measure change in performance over time.</p> <p>Compared to objective quality rating as well</p>	<p>Mean technical quality scores increased significantly as did those for quality of carotid patch. Inter-rater reliability for technical outcome (over all tutor impression) was assessed</p>	<p>Responsiveness of the PBA form was assessed. Reliability of the PBA form was found to be low, but could be related to the statistical analysis of the</p>

**Table 4 Continued**

	measures of change. Each PBA checklist item was given a score from 1 inadequate to 5 expert		and found to be acceptable at Cronbachs alpha 0.7. But surgical skill assessment as derived from the PBA form was low Cronbachs alpha 0.47 for task specific and 0.165 for global surgical skills	checklist being quite different to previous methodologies. Validity not addressed. Partial blinding only of the technical quality assessment.  Single centre
--	---------------------------------------------------------------------------------------------	--	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--------------------------------------------------------------------------------------------------------------------------------------------------------------------

Table 5 Example of the output from a “d study” on reliability of Procedure based assessment in the workplace

Cases	Assessors per case		
	1	2	3
1	0.69	0.75	0.77
2	0.85	0.88	0.89
3	0.91	0.92	0.93
4	0.93	0.94	0.94
5	0.95	0.95	0.96
6	0.96	0.96	0.96
7	0.96	0.97	0.97
8	0.97	0.97	0.97

5. Source: Taken form “Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology” (58).

This table illustrates that only 1 assessor for two cases is required to achieve reliability co-efficients above 0.8. Taken from the post hoc analysis section of the Health Technology report the original citation is table 57b after obstetrics and gynaecology trainees had been excluded from the analysis.

Since the publication of this study two further studies have been performed which have utilised PBA as an assessment tool. The first of these assessed 10 trainees performing a patch angioplasty on a pulsatile vascular simulator before and after simulated practice(122). PBA was used to assess the trainees at baseline and after a 3 day intensive training course. The PBA form was found to be internally responsive to a training effect but reliability of the checklist items was low. Validity of the PBA tool was not addressed in this study.

The final study described utilising PBA examined how accurate trainees were at performing self assessment on live appendicetomy compared to their supervisors(121). Twenty five higher surgical trainees took part and the main findings were that global summary scores correlated well between trainees and supervisors, with minor over-ratings on the checklist section of the PBA tool. Interestingly this study also included a qualitative assessment and identified that trainees were more likely to highlight non-technical skill deficiencies rather than technical errors, which supervisors tended to focus on.

Although the evidence base for PBA is presently small, it is clear that in well designed robust studies that it appears to be a valid and reliable tool in the theatre environments. However the evidence in the simulated setting is less robust although encouraging. Validity appears to be consistent when described in simulated studies, but there are conflicting results for reliability, possibly due to the different statistical analysis utilised to describe the results. Responsiveness, where assessed, is inferred rather than highlighted as a psychometric property of the PBA tool(97).

## **1.4 The case for Simulation: An Overview of the Evidence**

Simulation is defined as the ability to recreate authentic problems or scenarios to engage and train a learner in a particular skill, behaviour or characteristic that is divorced from real world consequences(31). In the context of medical and surgical education, simulation reduces the need for trainees and junior doctors to practice new skills on patients which could result in harm or distress and allows doctors the ability to repetitively and deliberately practice in order to develop expertise(37).

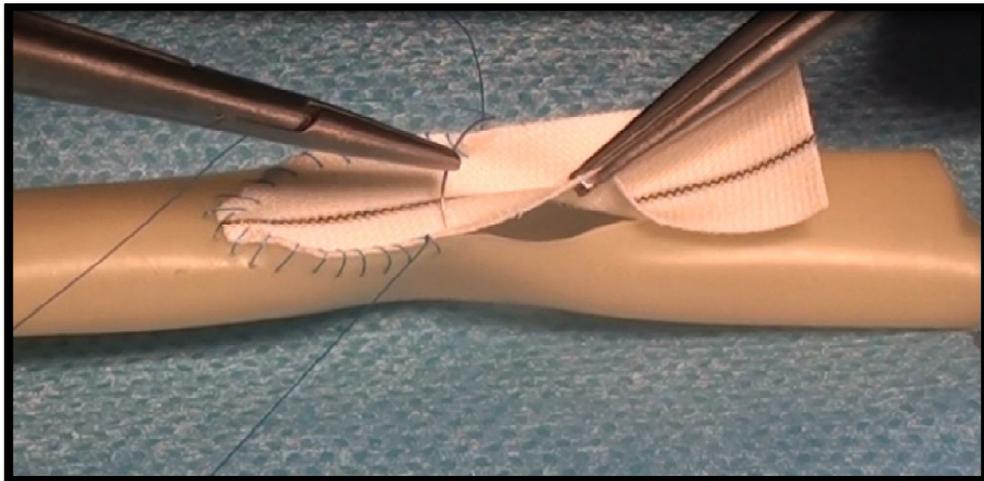
Simulation and simulation training encompasses a wide range of disciplines and methods of delivery (20, 123). Traditionally surgical apprenticeship models of training focused on practicing on patients through learning on the job and many years of observation of mastery before being allowed to perform surgery independently on patients(124), but development of simulation technologies is beginning to replace this traditional training paradigm, and is becoming increasingly realistic and complex as the technology in this field advances(125).

Simulation can broadly be classified into scenario training, low fidelity skills training, high fidelity skills training and perhaps the most complex which involves high fidelity skills simulators with real-time scenario training(125).

Scenario training encompasses simplistic one to one training such as patient- doctor interactions that can be found at a medical school which help students in techniques like breaking bad news or taking consent or counselling patients(126), to more complex recreations such as crisis management or advanced life support moulages. Here actors or other trainees and facilitators are used to represent different individuals to help trainees problem solve and practice complex technical skills such as intubation and resuscitation using high fidelity manikins(125).

Low fidelity skills training involves the use of training simulators that have less than life like handling properties and usually involves bench models such as plastic recreations of pelvic anatomy for practicing urinary catheterisation skills or plastic tubing or latex materials which can be used to practice vascular surgery techniques(127). There are also commercially available jig sets that mimic a part of an operation, such as a saphenous femoral junction ligation(92) or carotid endarterectomy(128) in vascular surgery. These types of simulator are low cost and often disposable but allow junior trainees the opportunity to practice basic techniques without consequence and with minimal cost. An example of a basic training surgical training jig is shown in figure 3.

**Figure 3 an example of a low fidelity vascular surgical training jig**



High fidelity simulators are relatively new and have been introduced in the last two decades. These types of simulator are designed to actually represent part or all of an operation. They frequently contain a computer and a mechanism by which a trainee can interact in a physical sense with the simulator. High fidelity simulators now exist for laparoscopic biliary(129), urology(130, 131) and gynaecological surgical operations. Endoscopic simulators are commercial available for colonoscopy, upper

gastrointestinal endoscopy and urology(132). Endovascular simulators have been designed to recreate interventional procedures such as angiogram and angioplasty for cardiologists, neurologists, radiologists and vascular specialists; they are typically expensive and have been limited to a small number of training units(133). Trainees and trainers require time to become familiar with the programme and systems that run the computer.

High fidelity simulation has not just been limited to the surgical disciplines and there are now resuscitation manikins that talk, breathe, groan, bite and have pulses and breath sounds which are increasingly being used to train doctors and allied professionals in life support techniques(134).

However systematic reviews show simulation is not superior to traditional training models such as learning in an operating theatre on real patients(123). But it has been shown to reduce error, shorten learning curves, reduce operating and procedural times and provide increased trainee confidence in performing complex procedures(26, 135). So, these facets of simulation; the ability to provide accelerated learning through deliberate practice outside the clinical environment are why there has been a growing emphasis in incorporating simulation into surgical curricula.

However there seems to be a lack of coherence within the literature of a structured approach to translating the evidence, which is abundant, from the small scale studies that presently exist in the surgical simulation literature to evidence based surgical simulation curricula design. The exception is presently the innovation of the pilot orthopaedic surgical training programme from Richard Reznicks' group in Canada, where a radical overhaul of surgical training is taking place; training surgeons in the simulation suite before being allowed to progress to operating on live patients(136-139). This structured modular approach aims to combine the benefits of simulation with specific directed learning opportunities in the operating theatre. The benefits of this approach are currently being

evaluated and tested but appear promising(137-139). In the first of these studies associated with this programme, 6, 1<sup>st</sup> year orthopaedic trainees on an intensive simulated skills competency programme were compared at baseline for technical skills with 8 on service and 8 off service trainees using OSATS as the mechanism of assessment. There were no differences at baseline between groups but after completion of the skills course all trainees were again assessed and those on the intensive course scored higher than the other two groups in performance of technical skills(139). This benefit was maintained at 7 months post intensive training and approached the competency levels of senior orthopaedic trainees(137). This study group have also demonstrated that improved performance can be achieved when directed student led practice verses a traditional demonstration approach, is incorporated into training programmes(138).

Apart from the trainee centred benefits that can easily be measured such as improved operative efficiency, self confidence and shortened learning curves simulation has a role to play in increasing patient safety(125, 135, 140). Simulation may minimise complications and errors that can occur through trainees operating on patients during the early phase of their training and many authors cite this as the main reason for introducing simulation into surgical training in a comprehensive manner. However proving such a benefit would be extremely difficult and as yet there are no large good quality longitudinal studies to suggest that simulation training reduces complications or errors in surgical performance in the operating room on patients. Further work needs to be directed at translating evidence of benefit from the simulation suite to the operating room.

However what has been established is the premise that deliberate practice develops expertise. Ericsson has written extensively on this subject and studied many experts in the fields of chess, music and professions such as airline pilots and physicians where complex motor skills and cognitive

patterns are required to be integrated on a daily basis(32). Deliberate practice is the hallmark of mastery and is required to progress a student beyond competence to the level of expertise. Traditionally in the field of surgery, deliberate practice was undertaken on patients, by surgeons in training and following post graduate training once surgeons achieved consultant status. Many people now suggest that this position is no longer justifiable where increased accountability is required and shorter training programmes have been introduced in the wake of the New Deal and EWTR(141).

Seeking new ways to train surgeons to a level that is considered competent, safe, and accountable and that can be delivered in the shorter duration of training that has now been advocated and also which will be relevant to the changing health needs of the population and the requirements of speciality training, will be challenging. The evidence is growing that simulation can play an important role in the content of surgical curricula in the future. How simulation training can be integrated and delivered within surgical curricula requires careful attention the evidence of relevance, opportunity, fidelity and assessment processes that will be involved.

In the following chapter the evidence base for commonly used vascular surgical simulation technologies and relevance to the experiments in this thesis will be explored and discussed.

#### **1.4.1 Bench Models**

Bench model simulation is perhaps one of the earliest forms of simulation training that has been integrated into surgical training in the past. Courses such as the Basic Surgical Skills (BSS) and Basic Laparoscopic skills course that have been traditionally run by the royal colleges in London and Edinburgh use this basic technology to train junior surgeon in basic

techniques such as instrument and tissue handling skills and suturing and knot tying skills. The Basic surgical skills course has been running since 1994 and provides junior surgical trainees a standardised approach to many surgical techniques.

Bench models include simple suturing jigs, plastic anatomical models and box trainers for basic laparoscopic work often made of foam or latex type materials. They are often considered low fidelity simulations due to less than life like handling properties and the limited nature of the skill being tested; often only part of a procedure or operation; however they have distinct advantages over practicing in the operating room. They are relatively inexpensive, reusable, require minimal or no supervision and are safe with no exposure to preservative chemical such as formalin or exposure to live animal or human tissue.

The earliest described efforts in modern day literature to introduce simulation training in to surgical curricular involving surgical low fidelity jigs is described by Bevan in 1981(142). He describes jury rigged simulators which allowed 18 surgical trainees perform gastrointestinal anastomosis on a 3 day structured course and since that time commercial simulators have become available and mass produced to allow structured teaching in basic techniques. However the literature describing benefit for such simulators is heterogeneous. There are a plethora of studies describing different simulators including vascular(36, 143, 144), gynaecological(145), urological(146), orthopaedic(97, 147), cardiac(116) and basic laparoscopic basic box trainers(145, 148). However the literature to evaluate their effectiveness is not coherent. Methods used include simple course evaluation questionnaires(149), global rating scales(148), checklist rating scales(145), and efficiency of motion analysis and video analysis(27, 72) which makes comparison difficult between similar studies. In addition many studies do not describe validity or reliability evidence and rarely both

and a typical study contains only small numbers of participants with varying levels of expertise (123, 127).

Sutherland et al in 2006 performed a systematic review of randomised controlled trials of simulation training(123). Within this study which identified 30 randomised controlled trials and 760 participants there were only 4 studies(149-152) that compared model training with other forms of training or no training and only 5 studies that compared laparoscopic low fidelity bench top box training with no previous training(151, 153, 154) (155, 156) The maximum number of participants with these studies was 24 which included both arms of the study(151). The outcomes between studies varied between objective assessment on a simulator(151), performance on cadaveric material(149) or performance in a live operation(152) or a translational benefit measure in the operative environment(150). The results were shown to be inconsistent to benefit with respect to box training and a modest benefit with respect to model training(123). But there are many confounding factors which are illustrated by such studies. Baseline characteristics vary from study to study as does the methods of evaluation and intervention, blinding is not always present and in some instances the studies do not state whether the simulation has been validated prior to the study being carried out, this makes interpreting the results difficult and in some instances meaningless, if the simulator being studied does not demonstrate a learning effect between groups of different ability then a ceiling effect is achieved and improvement between simulators or after repeated practice will not be demonstrated at the time of evaluation, will limit the usefulness of the simulator and may be difficult to prove translational benefit to the operating room.

In the area of vascular surgery replication of open surgical procedures has been demonstrated in a small number of studies. Pandey et al have

described the introduction of open surgical simulation in the introduction of the European Vascular workshop using 15 candidates of varying expertise and found using an abdominal aortic simulator improvements in global performance and time taken to perform and anastomosis were demonstrated from baseline after 3 days of training on an aortic simulator. Improvements were demonstrated using 3 blinded video raters for the pre and post assessment score using OSATS assessment tool(157). A further study by Pandey et al conducted over 2 years prior to the introduction of the skills assessment in the European Board of Surgery Qualification in Vascular Surgery, demonstrated that bench models of a sapheno-femoral junction ligation, synthetic tibial artery anastomosis and a knot tying simulator were construct valid and reliable methods of assessment and that the skills assessment did not correlate with oral examination performance or logbook accredited scores(158) which also illustrated the need for integration of psychomotor skills assessment in surgical training and accreditation .

Further single institution studies in assessing the usefulness of vascular bench models have had varying outcome measures; In 2003 Bann et al assessed 47 higher surgical trainees in Hong Kong in performing a vein patch at depth as well as a range of procedures, using a global OSATS rating form(159). This aim of this study was essentially a validation and feasibility study. While the trainees had received some training at a basic surgical skill level and had a course manual to refer to, the authors' report there was no correlation between level of training and the scores reported in this task, likely, they concluded to lack of vascular training within the surgical curriculum at that time.

Datta et al performed a similar study in 2005 using 22 surgeons(143). Five basic surgical trainees, 8 junior registrars, for senior registrars and five consultants performed a vascular anastomosis and sapheno-femoral

junction dissection which was assessed using an OSATS rating scale and motion analysis software. This was also a validation and feasibility study which demonstrated significant improvements with increasing seniority.

There are 4 randomised controlled trials which have looked specifically at vascular bench model procedures, again the outcome measure are variable between studies(92, 116, 160, 161). In 2007 Sidhu et al performed a single blind randomised trial using 27 surgical residents to assess the benefit of simulator fidelity on subsequent performance(160). Trainees were randomised to perform a graft to artery anastomosis either on a low fidelity plastic bench model or human cadaveric brachial artery in a 3 hours session. They then performed a graft to vessel anastomosis on a femoral artery in a live anaesthetised pig. Raters were blinded to which training method the surgeons had received. Both junior and senior trainees that had received training on the high fidelity cadaveric model performed better in the final product analysis. Suggesting that model fidelity also plays an important role in eventual performance.

Model fidelity can also be used to help discriminate between differing levels of competence, this is quite nicely illustrated by two studies by Black et al(128, 144) who first assessed junior and senior vascular trainees performing a simulated low fidelity carotid endarterectomy and compared it to the performance of consultants. There were no differences in the result between senior trainees and consultants, only junior and senior trainees, suggesting a ceiling effect for this simulation at senior registrar level(128). However in a later study he was able to demonstrate that the addition of a crisis scenario in a simulated theatre, adding an increased element of realism, with the same low fidelity model was then able to distinguish between senior trainees and consultants who performed better in the crisis scenario.(144)

The second randomised control trial assessed the role of deliberate practice in surgical performance. Price et al in 2011 published a single blinded RCT in 39 surgical trainees(116). Trainees were randomised to expert guided tutorial or expert guided tutorial in addition to self directed practice performing an end to end anastomosis on a commercially available synthetic jig. Self directed practice consisted of 10 vascular anastomosis performed over a 2 week period. Both cohorts of trainees were then assessed using an OSATS rating scale performing the same anastomosis on the carotid artery of a live anaesthetised pig by 2 blinded raters. Those trainees who undertook deliberate practice received higher OSATS scores, scored higher on end product evaluation and performed the anastomosis faster. The requirement for supervised deliberate practice was highlighted in a third RCT of simulated aortic repair where improved performance from baseline, as measured by OSATS was significant, when trainees were randomised to be supervised by trained faculty verses a lab coordinator(162).

Skills transfer to live operating performance is being increasingly recognised as an important discriminator in determining whether a simulation intervention is useful as a training tool. As, despite the positive evidence from studies such as orthopaedic group in Canada, training can be simulator specific, and these simulator skills need to be demonstrated to translate well to the real world. In the fourth of the randomised controlled studies in vascular simulation, Hseino and colleagues described the one of the first randomised controlled trials of skills transfer in vascular surgery to the live operating format using sapheno-femoral junction bench models(92). They randomised 12 procedure naive surgical trainees to bench model training or no training before asking them to perform a sapheno- femoral junction dissection and ligation in the operating room on a live patient while being assessed by a blinded rater. Those trainees that had received the bench model training outperformed the trainees who had

had no simulation training using an OSATS tool as measured by the global rating scale  $p < 0.001$  and procedure specific elements  $p < 0.001$ . While this is a small trial, with small numbers of trainees, it clearly demonstrates the benefits to the trainee in being given the opportunity for relevant, index specific practice before being allowed to operate on patients.

The main open vascular simulation studies are summarised in table 6, (pages 75-77) but of the evidence that presently exists for vascular simulation there is emerging data to suggest that bench models can improve performance after deliberate practice and assessment, that higher model fidelity improves retention of skills and can further discriminate between skill level of participants and that practice on simple bench models can improve performance in the operating room, when assessed by validated methodologies.

**Table 6 Summary of the main vascular simulation studies involving bench models and similar techniques**

Study	Setting	Outcomes	Comments
<b>Pandey et al 2005(157)</b>	15 candidates European vascular workshop using aortic abdominal simulator were assessed at baseline and after 3 days training	Global performance and time taken to perform anastomosis using blinded video raters utilising the OSATS tool.	Construct validity of the simulator or OSATS tool not explored, simply the training potential of the simulator, and illustrating that the assessment tool could be responsive
<b>Pandey et al 2006(90)</b>	30 candidates of different training level tested on sapheno-femoral junction, distal tibial anastomosis and knot tying simulators	Poor correlation with exam performance and log book accredited scores but demonstrated to be valid and reliable methods of assessment (intra-observer reliability)	Assesses only the difference in performance between assessors and candidates but highlighted the present lack of psychomotor assessment in surgical training before consultant certification
<b>Bann et al 2003(159)</b>	47 higher surgical training were assessed performing a vein patch at depth, along with additional open procedures.	Main aims were validation and feasibility study. Vein patch insertion was found to be not valid assessed by OSATS form.	Vein patch insertion found to be unreliable due to lack of vascular training in that particular curriculum.  Single institution, un-blinded
<b>Datta et al 2006(143)</b>	22 surgeons, five basic trainees, 8 junior registrars and 4 senior registrars and 5 consultants performed a vascular anastomosis and sapheno-femoral junction dissection	Assessed using motion analysis and OSATS assessment forms. Which demonstrated increasing scores with seniority and so demonstrated validity of these models	No assessment of reliability, but OSATS scores correlated well with motion analysis
<b>Sidhu 2007(160)</b>	27 surgical residents were randomised to performing a vascular anastomosis on a plastic bench model or human cadaveric on a brachial artery in a 3 hour session. Residents were then assessed on a live pig model performing a graft to femoral artery anastomosis by raters blinded to training technique	To assess the impact of model fidelity on performance.  Construct validity demonstrated for junior and senior residents for the final femoral artery anastomosis	Validity of the training models not discussed.  Reliability not assessed by blinded final assessments.  Demonstrated that Model fidelity plays an important role in skills transfer and Authors found that cadaveric training conferred the greatest benefit.

**Table 6 Continued**

<p><b>Price et al 2011(116)</b></p>	<p>39 trainees randomised to expert guided tutorial or tutorial plus self directed practice of 10 end-to-end anastomosis on a synthetic jig. Then assessed on a carotid artery of a live anaesthetised pig using OSATS</p>	<p>To demonstrate that self directed deliberate practice improve performance.</p> <p>OSATS scores, final product analysis and time to complete the procedure were all higher in the deliberate practice group</p>	<p>Demonstrated that expert guidance improved scores.</p>
<p><b>Hseino 2012(92)</b></p>	<p>12 procedure naive trainees randomised to sapheno-femoral junction (SFJ) ligation on a bench model or no training. Trainees then assessed by a blinded rater using a OSATS tool and ICEPS for sapheno-femoral junction ligation on a live SFJ operation</p>	<p>Designed to assess skills transfer</p> <p>All trainees assigned to simulator training out performed those that received no training on both the OSATS and ICEPS tools</p>	<p>Demonstrated responsive nature of the assessment tools used and that skill transfer at Sapheno-femoral junction ligation was feasible.</p>
<p><b>Black 2007(128)</b></p>	<p>41 surgeons, (consultants, senior and junior trainees) assessed by validated rating scales on a carotid endarterectomy model</p>	<p>Validity and reliability assessed.</p> <p>Significant differences were found between junior and senior trainees, but not senior trainees and consultants, suggesting partial validity of this training model, it was found to be highly reliable Cronbachs alpha 0.9</p>	<p>Validity and inter-rater reliability both assessed.</p> <p>Model did not demonstrate full validity between senior trainees and consultants, suggesting a ceiling effect for this simulation.</p>
<p><b>Black 2010(144)</b></p>	<p>High fidelity carotid endarterectomy in simulated operating room environment with 30 surgeons (10 junior trainees, 10 senior trainees and 10 consultants). Assessed using a validated rating scale and non technical skill assessment in a crisis and non crisis scenario</p>	<p>Validity and reliability assessed</p> <p>Assessed technical and non technical skills and confirm validity for all levels of participant for both crisis and non crisis scenarios and inter rater reliability was high &gt;0.8 for both scenarios</p>	<p>Versatile assessment in a high fidelity environment.</p> <p>Validity and reliability of the scenario assessed.</p> <p>Added high fidelity provided a discriminator between senior trainees and consultant which was not seen in the previous low fidelity study.</p>

**Table 6 Continued**

<p><b>Robinson 2013(162)</b></p>	<p>18 surgical residents randomised to 2 sessions of simulated aortic aneurysm repair either supervised by trained faculty or skill lab co-ordinator. Used a validated tool similar to the OSATS</p>	<p>To assess improvements from base line in both cohorts. Only the group assigned to trained faculty improved significantly from base line</p> <p>Modified OSATS found to be valid for year of training at baseline, and the simulator was found to be valid at baseline for year of training.</p>	<p>All assessments blinded and assessed by video.</p> <p>No reliability analysis.</p>
<p><b>Duran 2014(115)</b></p>	<p>Ninety two trainee assessed performing a end-to side anastomosis using a modified OSATS tool</p>	<p>Outcome primary to assess the validity and reliability of the OSATS tool. Found to be valid between novice, intermediate and senior trainees, for the checklist and global rating scale GRS. Inter-rater reliability ranged from 0.64-0.77. high internal consistency between checklist and global rating scale high.</p>	<p>Study methods abandoned the checklist element as cumbersome and redundant given high internal consistency between checklist and GRS.</p> <p>Validity, reliability and internal consistency addressed.</p>
<p><b>Robinson 2012(161)</b></p>	<p>37 junior residents randomized to a 3 week course or 6 week course of instrument recognition and end-to –side vascular anastomosis. Assessed using a modified OSATS checklist score converted to a numeric value.</p>	<p>To assess improvements from baseline between a short or long course.</p> <p>All trainees improved from baseline and there were no significant differences between short and long course. Skill retention at 16 week was similar</p>	<p>Validity of the assessment tool not described, nor reliability.</p> <p>The tool appears responsive to practice effect.</p> <p>Clear benefit in skill improvement</p> <p>Blinded only to allocation</p>

## 1.4.2 Endovascular Simulation

The first endovascular simulator to be described was by Dawson et al in 2000(163). This essentially consisted of a computer based synthetic fluoroscopy unit, interactive anatomical display and enabled the user to perform left and right coronary artery catheterisation using real catheters. There are now 3 other endovascular simulators that are now commercially available; The CathLabVR surgical simulator (Accutouch® System Immersion Medical), the ProCedicus Vascular Intervention Simulator Trainer (VIST™ Mentice) and the ANGIOMentor™ family (Simbionix)(133). Each of these simulators has simulated fluoroscopy and provides the user with tactile or haptic feedback and a greater or lesser degree of simulated patient monitoring. At the end of each procedure the user is given a detailed breakdown of performance using simulated metrics such as time to complete procedure, contrast and fluoroscopy use and depending on the procedure performed, a quality of final product analysis such as residual stenosis in the case of angioplasty or stenting.

Due to the multiple specialities that can potentially benefit from the training possibilities of virtual reality high fidelity endovascular simulation, such as vascular surgeons, neurosurgeons, cardiologists and interventional radiologists the European Virtual Reality Endovascular Research Team (EVEREST) was formed. This was mainly fuelled in the first part by the need for training in carotid artery stenting (CAS)(133). This is a high risk procedure with a distinct learning curve and serious potential complications (risk of stroke to be between 5-6% as opposed to carotid endarterectomy with a stroke risk of 1-2%) and the high fidelity simulators which allow rehearsal of this complex procedure were felt to be ideal in order to shorten the learning curve associated with this procedure,

therefore much of the early work of endovascular simulation focused on CAS. Since 2004 there have been a number of studies focusing on renal artery angioplasty and stenting and iliac and superficial femoral artery (SFA) angioplasty and stenting in addition to CAS(110, 164, 165) and are summarised in table 7 (page 84).

In common with the work regarding open simulation these studies typically involve a small number of candidates and the outcomes measured are variable. Much of the early work performed on endovascular simulators involved experienced interventionalists performing CAS. Dayal et al in 2004 reported on the results benefits of using VIST simulator to perform CAS(166). Twenty one interventionalists (5 experienced interventionalists with >300 procedures and 16 novices with <5 interventional procedures) performed two CAS procedures after didactic teaching and were assessed using a 50 point checklist scale and simulator metrics. They were able to demonstrate statistically significant improvements in novices' procedure time, fluoroscopy use and checklist scores which was not demonstrated in the experienced interventionalist group. In essence this was a construct validation study which demonstrated a learning benefit in inexperienced trainees. Hsu et al in 2004 also published a similar study on CAS using the VIST simulator, here they assessed 41 subjects, novice and advanced cohorts performing CAS(167). Only 29 completed the study protocol which consisted of a single pre-test run on the simulator performing a CAS followed by either further practice over 60 minutes or no practice. Those that received practice improved their completion time significantly and the greatest effect was seen in novice trainees confirming that the simulator was construct valid for CAS and had the potential for training benefit.

Over time the versatility of endovascular simulators has been explored and further studies have been designed to assess the benefit of endovascular simulators in renal and iliac and SFA interventions. In 2006 Berry et al

looked at simulator metrics alone in a study group of 8 experienced interventionalists and 8 novices (medical students) performing a renal artery stenting procedure(168). Each participant received 45 minutes of didactic teaching followed by a 2 hour familiarisation session then a testing period. There were no significant differences between groups in residual stenosis, stent placement accuracy or number of cine loops performed or procedure time. The only significant difference was in fluoroscopy use which was greater in the novice group, suggesting lack of construct validity in this method of assessment and simulation, but the educational benefit of the simulator for the novice trainees was not emphasised.

Neequaye et al in 2007 assessed the benefits of skills transfer from renal stenting procedures to iliac stenting(169). They randomised 20 trainees without endovascular experience to perform either eight renal artery angioplasties or iliac artery angioplasty using the VIST simulator. Each trainee then crossed over to perform the alternate procedure twice. Their performance was assessed using the metrics provided by the simulator. The main findings of this study demonstrated that there were demonstrable learning curves for both procedures in time taken to complete, fluoroscopy, stent placement accuracy and contrast use. However those trainees that had performed the iliac intervention first had significantly higher fluoroscopy times when performing the renal intervention than those that had performed this first. Whereas those that had performed the renal intervention first performed the iliac intervention to the same level as the initial iliac group, suggesting that selective arterial catheterisation is a more complex and separate skill.

There are no validation studies to date that have addressed the suitability of iliac or SFA angioplasty as a training tool in surgical or interventional trainees. This seems rather counter intuitive to a staged curriculum approach that would be advocated in the clinical setting. For example any

interventional trainee would not begin learning interventional procedures by performing a CAS or renal intervention as these represent more complex procedures that would be undertaken by experienced interventionalists as perhaps demonstrated in the above study. Interestingly those studies that do involve either iliac or SFA angioplasty are the studies that illustrate an in vivo benefit or transfer of benefit to the clinical setting. In 2007 Berry et al published the results of a small study which stratified, by experience, 12 surgeons and interventional radiologists (open surgical experience 0-31 years and endovascular experience 0-5 years) to 4 groups which were allocated two training sessions(111). This was in either iliac artery revascularisation in a porcine model or in a virtual reality simulator or a combination of both in either order consisting of repeated practice over 3 hours. They then completed a further evaluation in both models. They were able to demonstrate using blinded video ratings of both procedures that exposure in the virtual reality model improved performance in both the virtual reality model and the porcine model. Practice on the porcine model improved the final porcine model score but not the virtual reality model score suggesting once again that model fidelity plays an important role. Virtual reality model scores were consistently higher than the porcine model scores and surgical or interventional experience did not appear to predict total scores although total scores for experience in both models combined improved significantly. Analysis of median total scores suggests no ceiling effect was reached in either model. This study suggest that virtual reality training can be translated to an in vivo setting, but this study contained small numbers of experienced practitioners , 3 per group, and lacked any prior validation of either model or their assessment tools, a modified OSATS tool and checklist. This may explain why no differences were found between differing levels of experience; despite the authors statement that total scores skills

assessment is independent of the “laborious necessity of validation of simulator metrics”(110, 111).

Chaer et al in 2006 is the only randomised, but un-validated, trial published to this present date regarding the translation benefits of endovascular simulation to the catheter lab on real patients. This study describes a two arm study in which 20 interventional naive trainees received either didactic teaching on interventional catheter techniques or didactic teaching combined with a maximum of 2 hours teaching on an iliofemoral angioplasty/stenting model on the VIST simulator. Randomisation was based on sealed envelopes after all trainees had completed the didactic teaching stage. Each trainee then went on to complete 2 lower limb interventions within 2 weeks of completion of the training by a blinded assessor using a checklist and global rating scale which was adapted from the original OSATS tool. The main finding of this study were that those trainees which received prior simulator training performed better in the majority of steps and global rating score when required to perform an intervention on real patients than those that did not, an effect that persisted in the second intervention, interestingly the most significant values are seen in the elements of wire handling and catheter handling and precision of wire and catheter techniques and general flow of the procedure, skills not assessed by the simulator directly. There were no validation processes described within this study prior to the trainees performing the iliofemoral intervention or if the adapted checklist and global rating scale described was a valid measure of performance.

A further prospective study is presently in progress in the United States which has been running for 4 years and involved 25 surgical residents which have been randomised to intensive simulator training or a standard surgical rotation which has not yet been published, Lee et al(170). Early results have shown an additional average of 9.3 hours of training on a

simulator has resulted in a translational benefit in live case performance (scores of 3.6 vs. 2.4 out of 5  $p < 0.001$  as measured on a Likert scale). All residents which underwent simulation training felt it helped with learning new skills in the angiography suite and improved their 3D conceptualisation and their understanding of imaging techniques. As it has not yet been published it is difficult to comment on the methodology of the study but it illustrates the potential of endovascular simulation as a training tool for the future, and this and studies such as Chaer et al, illustrate the potential benefits of simulated endovascular training for vascular specialist trainees.

**Table 7 Summary of the main endovascular simulation studies illustrating the simulator type and procedure investigated**

Study	Device	Module	Construct validity	Transfer to in vivo
Dayal et al 2004(166)	VIST	Carotid	Yes	No
Hsu et all 2004(167)	VIST	Carotid	Yes	No
Nicholson et 2005(171)	VIST	Carotid	No	No
Aggarwal et al 2006(172)	VIST	Renal	Yes	No
Hislop et 2006(112)	VIST	Carotid	No	No
Berry et al 2006(168)	VIST	Renal	Yes	No
Patel et 2006(171)	VIST	Carotid	No	No
Chaer et al 2006(110)	VIST	Iliac/SFA	No	Yes- Human
Dawson et al 2007(165)	Simsuite	Iliac	No	No
Berry et al 2007(111)	VIST	Iliac	No	Yes- Porcine
Neequaye et al 2007(169)	VIST	Iliac/renal	No	No
Van Herzeel et al 2007(173)	VIST	Carotid	Yes	No
Van Herzeel et al 2008(174)	ANGIOmentor	Carotid	No	No
Lee J.T 2012(175)	ANGIOmentor	Iliac and renal	No	Yes- live performance

### **1.4.3 Video Analysis**

Video taping of surgical procedures has a number of roles within the educational literature. It can be used as an educational tool and many websites and institutions such as the royal college of surgeons use videos of surgical procedures and simulations to provide a standardised procedure for reference in order help trainee surgeons learn new procedures and skills. One of the most valuable uses for the trainee is the potential it has to provide feedback on performance and there are studies that focus on the usefulness of video performance and its role in giving trainee feedback(38, 176, 177).

Review of the literature in vascular simulation suggests that video analysis has a limited role in objective assessment of trainees on a daily basis. The majority of studies within this area use video analysis as a tool to ensure blinding of rater assessment and to allow reliability data to be gathered. This usually occurs when either an assessment process such as a new simulation model is being tested or when a new tool for assessment such as PBA, OSATS or mechanisms that measure surgical efficiency need to be evaluated. As procedures become more complex and skills such as decision making are included in such studies, the limitations in video feedback and assessment become obvious. Few institutions have dedicated facilities to produce multiple camera angles that allow a comprehensive assessment of surgical performance which encompasses both the skill and required cognitive elements essential in advanced surgical practice that a live contemporaneous rater can provide. This of course in many situations removes the blinding element crucial for many studies, but it is not impossible, merely time consuming and difficult to produce under study conditions. As such this type of video assessment process is often limited

to small numbers within each study and often limited to giving feedback alone specifically on the technical skills being demonstrated. Under study conditions finding qualified raters, usually experienced clinicians, for whom it is often difficult to find time to evaluate a performance is also an issue. Each video must also be edited in a standardised fashion which also takes a considerable amount of time, even for short procedures. Once this is magnified up for each and every participant within a study the logistics can be easily become unmanageable.

The majority of studies using video analysis to assess trainees are therefore limited to single camera assessments using bench model simulation or open surgical procedures or laparoscopic tasks where the in situ camera makes assessment of psychomotor skills easier(96, 104).

Within the arena of vascular and endovascular surgery there are few studies that have used video assessment as part of the study protocol. The outcome measures include establishment of validity and reliability of assessment methods(104), establishment of construct validity of a simulator model(128, 143, 144), assessment of improved performance(38, 157) and assessment of construct validity of a endovascular simulation model and reliability of a modified OSATS tool(112).

Datta et al performed two studies in 2006. The first used video analysis of 2 previously validated surgical models of a small bowel anastomosis and vein patch insertion to assess the validity and reliability of snap shot video assessments scored with OSATS in 30 trainees(104). This was in comparison to a surgical efficiency score which was a combination of hand motion analysis and final product analysis and in comparison to the entire procedure which was videoed. The role here was to utilise the video assessment in order to provide blinded video raters to determine which form of assessment correlated most highly with the gold standard assessment method of OSATS, in this case the authors found that the

surgical efficiency score was the most reliable assessment method and correlated highly with the gold standard OSATS appraisal method. The second study as described by Datta et al in 2006 utilised blinded video assessment to determine the construct validity of vascular anastomosis and sapheno-femoral junction bench model dissections by 22 trainees and consultants which demonstrated significant differences between junior and senior trainees as measured by the previous validated OSATS assessment tool, thereby confirming that these particular models are construct valid as measured by OSATS(143).

Black et al published a similar study in 2007 in which they utilised blinded video analysis of 41 surgeons performing a CEA procedure on a bench model and a three throw knot tying exercise to determine if these particular bench models could distinguish between different levels of surgical expertise or construct validity(128). In this particular case, there were significant differences between junior and senior trainees, but not senior trainees and consultant which are a similar finding to Datta et al in their earlier study. They too used a generic OSATS rating scale and a final product analysis score.

Demonstration of improved performance after training using video analysis is in common to the studies by Pandey(157), Boyle(38) and Sigounas(178). In the first of these studies by Pandey et al video analysis was used at the European vascular workshop in 2003 and 2004 and involved 15 consecutive participants performing a proximal anastomosis on an aortic simulator before and after intensive training. Assessment methods include time, quality of final product and generic and procedural rating scale (OSATS). All participants' demonstrated improvement however retrospective analysis of the results suggested those that did not improve was those with the most prior experience(157).

Boyle et al used video analysis of handling errors of endovascular instruments in combination with simulator metrics in 18 endovascular novices performing renal artery stenting 6 times on an endovascular simulator. This study was able to demonstrate that trainees who received expert feedback improved the most in comparison to non expert or no feedback as demonstrated by simulator metrics and video assessed handling errors. This study illustrates a significant finding on the literature on endovascular simulation which is the inability of the endovascular simulator to provide relevant feedback to the trainee and the limitation of simulator metrics to distinguish between relevant or irrelevant handling errors which can often be over estimated by the simulator(38). This concept was explored further by Hislop and colleges who used blinded video assessment by 2 raters of 61 subjects performing three vessel wire and catheter advancement of the aortic arch. In an effort to capture all relevant information in a blinded manner they videoed each subjects hands performing guide wire manipulation and the internal fluoroscopic recording from the VIST endovascular simulator. They were able to demonstrate that the simulator was able to distinguish between inexperienced and experienced operators in completion times i.e. valid and in previous experience as measured by a modified OSATS scale. This suggests that the simulator is a valid construct in determining between different levels of expertise and that a modified generic OSATS was also valid for level of expertise. Reliability was also reported as statistically significant between the 2 raters, although no reliability co-efficient was quoted in this study(112). A summary of these studies is demonstrated in table 8 (p 89) but highlights the scanty literature on video analysis in simulated vascular surgery and associated techniques.

**Table 8 Summary of the main vascular simulation studies utilising video assessment method**

Study	Procedure Assessed	Methodology	Validity	Reliability
<b>Datta et al 2006(104)</b>	Vein patch	Snap shot edited video assessments scored with OSATS in 30 trainee vs motion analysis vs entirety of the video	Concurrent validity only with OSATS and the surgical efficiency score but construct validity not assessed	no
<b>Datta et al 2006(143)</b>	Vascular anastomosis and sapheno-femoral junction models	22 trainees and consultants assessed by blinded video using OSATS tool	Construct validity confirmed, between junior and senior trainees	no
<b>Black et al 2007(128)</b>	CEA and 3 throw knot tying exercise	Blinded video analysis using generic OSATS tool	Construct valid between junior and senior trainees but not senior trainees and consultants	No
<b>Pandey et al 2005(157)</b>	Proximal anastomosis on an aortic simulator	15 consecutive participants videoed performing before and after training assessed using OSATS	No validity or reliability assessment performed. Improvement of base line scores only	no
<b>Boyle et al 2011(38)</b>	Renal artery stenting	18 novices performing 6 stents, 2 groups with expert and no feedback. videoed	No validity or reliability assessed , study to highlight the importance of expert feedback which improved performance in this study.	no
<b>Sigounas et al 2012(178)</b>	Femoral to popliteal bypass, carotid endarterectomy and open aortic aneurysm repair	Senior general surgical residents, were video pre and post performance and assessed by a blinded video rater	To assess improvement in technique with dedicated practice. No validity	No reliability
<b>Hislop et al 2006(112)</b>	Three vessel wire and catheter advancement over the aortic arch	61 subjects, and 2 raters. Video assessment of both fluoroscopy and subjects hands, internal metrics and OSATS tool used	Simulator was valid for completion times, ie more experienced operators were quicker and received higher scores	Reliability was assessed but no co-efficient given.

## 1.5 Chapter Summary

Surgical training has undergone many reforms in the past few decades. Changes have taken place in the structure of the training pathways and working conditions for all surgeons under government and European legislation with the aim of improving working conditions for many hospital based speciality doctors. However this has been at the expense of less “on the job” exposure to training opportunities for junior trainees. Surgeons and many other “craft” based specialties have been affected most(16) where psychomotor skills form a significant part of the competencies that are expected to be achieved prior to completion of specialist training. In conjunction with these restrictions medical education as a speciality in its own right has driven changes in the methods used to assess trainees, especially in the area of surgery, resulting in unique behaviourally defined assessment methods mapped to clearly defined curricula.

The need for competency assessment and less exposure to live operating lists and theatre based learning opportunities have highlighted the need to supplement traditional training methods of learning on patients to perhaps more ethically, educationally sound and potentially more cost effective methods such as simulation training.

Basic low fidelity simulation training methods have been in standard practice for the past few decades, but only on an ad-hoc course basis, delivered intermittently to trainees at the discretion of deanery funding and study leave. But as technology in this area has improved, so has the potential to integrate simulation into many facets of surgical training on a regular, training level appropriate manner. In the discipline of vascular surgery there have been few open and even fewer endovascular studies that can help inform how we deliver and assess simulation training if it is to

be used as an adjunct in the future. Those studies that do exist which highlight the positive element of that deliberate practice improves performance are frequently, small, un-blinded and use assessment technologies that are either unfeasible in the workplace, lack validation or reliability data and few demonstrate evidence of transfer of skills to where it will matter most, in the operating room on live patients.

In order to coherently assess simulation training within the present curriculum, comparative data on the present method of assessing surgical trainees, PBA, needs to be sought. Although demonstrated to be valid and reliable in the operating room a question still remains as to whether there is an equivalent effect in assessing simulated procedures.

Therefore utilising a similar methodology to the largest study on PBA in the workplace to date(58) evidence was sought to demonstrate if PBA could be a valid, reliable and responsive tool in evaluating simulated vascular procedures. Specifically utilising the present evidence base where evidence for simulated aortic repair(157), sapheno-femoral junction ligation have been shown to improve trainee performance(90, 92, 158), evidence was sought to determine if PBA was valid and reliable method of assessing these simulated procedures and compare this assessment method to a previously validated simulation assessment tool. Given the present lack of endovascular evidence regarding entry level endovascular work in vascular novice trainees, particular iliac and superficial femoral artery (SFA) angioplasty a trainee based study was designed to evaluate the training benefits of an endovascular simulator utilising simulator metrics and to determine if PBA could be used to assess this method of simulation and if PBA could be proved to be a reliable and responsive assessment tool.

## **Chapter 2: Materials and Methods**

### **2.1 Chapter Overview**

The following chapter will describe the sequential experiments designed to investigate the psychometric properties of the assessment instrument PBA. Each experiment has been designed to investigate a specific psychometric element or elements of PBA.

Experiment 1 will explore the validity and inter-rater reliability of PBA in simulated bench top procedures, specifically, saphano-femoral junction ligation and proximal aortic anastomosis. Experiment 2 utilised a simulated endovascular procedure, a superficial femoral artery (SFA) angioplasty. The construct validity of the simulator had not been demonstrated before for this particular procedure and so prior to gathering data to assess inter-rater reliability and responsiveness of the PBA tool, a construct validity study was performed and the main findings will be illustrated in appropriate results section. Finally experiment 3 will illustrate whether PBA possesses intra-rater reliability in addition to inter-rater reliability and this will be demonstrated using a standardised video assessed by the same raters at two different time points.

### **2.2 Data Collection Methods**

The first of these tools, Procedure Based Assessment or PBA, is the operative assessment method in regular use for UK general surgical trainees in the workplace. There is minimal data on its use in simulated procedures(96, 97). In addition to the PBA tool, data on the simulated procedure was gathered using the generic section of the Objective Structured Assessment of Technical Skills or OSATS tool of which there is an abundance of data in simulated bench top procedures(58, 102, 107, 108)

and live operative assessments(58, 100). The inclusion of the OSATS tool allowed comparison to a known gold standard assessment method, in particular the generic assessment section, which has been proven to be valid and reliable in assessing simulated procedures.

For the purposes of this experiment each index specific PBA for sapheno-femoral junction ligation and aortic aneurysm repair and superficial femoral artery angioplasty were rationalised to include only those steps that may be seen in the simulated procedure. In the main part these consisted of the task specific and global elements of the competencies and definitions that can be found in section IV (exposure and closure) and section V (intra-operative technique) of the index specific PBA form. In addition the global summary score that is included in each PBA was also included for each assessment. This consists of a single generic score from 0-4 and corresponds to an anchoring explanatory statement of a trainees' performance and is summarised as follows;

- 0- Insufficient evidence to support a summary judgement**
- 1- Unable to perform the procedure, or part observed under supervision**
- 2- Able to perform the procedure, or part observed under supervision**
- 3- Able to perform the procedure or part observed with minimum supervision (occasional help needed)**
- 4- Competent to perform the procedure unsupervised (could deal with complications that arose)**

Each assessor then rated the trainee as described in the original PBA form. Where each competency was observed it was rated as N (not seen), D

(development required) or S (satisfactory for certification of completion of training). Each trainee was also rated on the single generic score from 0-4 (appendix 5).

The OSATS assessment tool as it was originally formulated consisted of a checklist of items followed by a generic score. Studies of this tool suggest that the most reliable elements of this assessment tool were the generic elements(102, 107) and therefore this generic element was included in this study to allow comparison to this gold standard. In essence this scoring system includes 7 domains, respect for tissues, efficiency of motion, instrument handling, use of assistants, knowledge of the instruments, and knowledge of specific procedure and flow of operation. Each domain is evaluated on a Likert scale from 1-5 anchored with explicit descriptors to allow the assessor to rate the trainee accurately. (Appendix 3) Therefore each trainee was given a single score out of 35. For the superficial femoral artery angioplasty the OSATS tool was modified from previous studies utilising this form in endovascular procedures and approved by the lead consultant interventional radiologist involved with the study and consisted of 8 domains therefore each trainee was given a score out of 40. An example of this form is given in appendix 6

For the purposes of this study both these assessment tools were amalgamated to a single, double sided sheet that could easily be used and distributed to the assessors and trainees for completion. In addition to both these assessment tools the form contained sections for the trainee and/or assessor to complete for grade of trainee, grade of assessor, number of times the procedure had been observed and performed before.

To determine construct validity of the simulator in experiment 2 data metrics produced by the simulator were gathered for all participants within the study, novice, intermediate and expert to determine which if any of these variables could prove the simulator to be a valid training tool and

construct. The data metrics produced by the simulator included time taken to complete the procedure, fluoroscopy, contrast use and residual stenosis. Additionally the simulator records the time taken to position the sheath for the intervention, number of catheters and wires used and number of contrast runs. In order to estimate reliability on the 5<sup>th</sup> and 10<sup>th</sup> run the novices were assessed using a rationalised PBA tool on angiogram and angioplasty modified from the Vascular Society website to reflect a simulated angioplasty (see appendix 5).

### **2.3 Statistical analysis**

In order to compare validity and reliability and responsiveness data across a range of different procedures a single consistent method was sought to analyse the categorical data generated from the PBA tool. Therefore after discussion with a statistician it was felt appropriate to analyse the data in line with methodology of the single largest study to date regarding PBA(58). This method converts the categorical data in the itemised competencies section of the PBA to a numeric value for each of the ratings. In a slight deviation from the method previously described by Beard et al(58), N=0, D=1 and S=2 was used rather than no N and D = 0 and S= 1 as described in the previous study. Using this method it was hoped that analysis of the itemised competencies would be more sensitive to change and the varying skill levels demonstrated. In order to allow comparison of these scores between different procedures the total score for each of the itemised competencies for the PBA was converted to a perfect score of 1 for each procedure and trainees were then allocated a score based on a ratio of their performance compared to the perfect score of 1. The single generic summary score from the PBA and generic OSATS score was analysed as a numeric value without modification. Where consistently competencies were not seen in a particular simulated procedure, such as “dealt with variations in anatomy” these competencies were omitted from

the final analysis consistently across procedures as being redundant to the analysis.

The Statistics Package for Social Science (SPSS 19) (Chicago, ill) was utilised to analyse the data. Due to the small number of data points within each study and after analysis of the frequency histograms of the data it was felt that non-parametric tests were the most appropriate to use. This allows less reliance on statistical assumptions of normality due to the non gaussian distribution of data(179).

**Validity-** To prove the directionality of the hypothesis stated 3 statistical tests were used.

- **Spearman's Rank correlation co-efficient-** to analyse non parametric continuous and ordinal data
- **Kruskal-Wallis one way analysis of variance** - for differences between non parametric multiple, unrelated, unequal sized groups (specifically construct validity of the endovascular simulator)
- **One Sample T-test** – In the context of determining levels of agreement between two measurements to assess concurrent validity(180).

**Reliability-** To determine the inter-rater and intra-rater reliability based on classical test theory 2 specific statistical tests were used.

- **Spearman's Correlation co-efficient** – to determine if a statistically significant correlation can be ascertained between raters
- **Cronbachs alpha-** as a hypothesis free estimate of reliability

Generalizable theory would have represented the ideal estimation of reliability however a sufficiently large data set was not available and the numbers of raters were limited to allow this form of statistical modelling.

**Responsiveness** – To determine if a statistical difference exists between the paired samples taken over time in a non-gaussian data set

- **Wilcoxon signed rank test** - to determine the paired difference between mean ranks in the same population.

For the purposes of data interpretation p-values are quoted to 3 decimal places and p-values <0.05 are deemed to be statistically significant and lead to rejection of null hypothesis of no difference between groups or no correlation of data where applicable.

## **2.4 Inclusion and exclusion criteria**

Inclusion criteria were:

- Willingness to participant in the study and complete the study protocols
- Ability to understand and use the modified PBA form

Exclusion criteria:

- Unwillingness to participate in the study
- Not familiar with or never used the PBA form before

## **2.5. Data Security and Confidentially**

All data was appropriately anonymised and stored either on hard copy in a locked research office or an encrypted password protected computer for statistical analysis. All individuals were assigned specific alphanumeric identifiers to refer to individual data points or raters.

## **2.5 Experiment 1- Validity and Inter- Rater Reliability of Procedure Based Assessment in Simulated Vascular Procedures**

### **2.5.1 Study Setting for Experiment 1**

The initial experiment designed to explore the validity and reliability of a modified PBA in simulated vascular procedures was performed at the Royal College of Surgeons England during two simulated training courses. The two courses involved were the Speciality Skills in Vascular surgery and Advanced Speciality Skills in Vascular surgery. This allowed an observational study to be performed on many trainees performing a variety of simulated procedures relevant to vascular surgery across a range of different levels of surgical trainee. These two courses also provided a pool of experienced assessors in vascular surgery who had little or no prior knowledge of the trainees being assessed.

### **2.5.2 Study participants**

The participants in this study included basic and higher surgical trainees from the United Kingdom and overseas. The training grades assessed ranged from speciality trainees ST1 and ST2, to the higher surgical training grades ST3 to ST5 and consultant. Where trainees admitted to non UK standard training grade they were questioned further to determine their number of years in surgical training and allocated to the equivalent UK training grade for the purposes of data collection.

### **2.5.3 Study Assessors**

In the first instance study assessors included the experienced vascular consultants and tutors who agreed to participate in the study and the ST4 speciality trainee coordinating the study. As many of the simulated stations required two trainees per station each trainee also evaluated the other in addition to the vascular consultant supervising the trainee. This increased

the number of assessments performed and allowed for inter-rater reliability data to be gathered for each of the procedures assessed.

#### **2.5.4 Simulated Procedures and Materials**

Two simulated procedures were assessed to gather validity and reliability data on the modified PBA assessment tool. The simulated procedures included a sapheno-femoral junction ligation and proximal aortic anastomoses provided by Limbs and Things™ and are low cost, low fidelity synthetic simulators, designed to approximate a specific step or section of a vascular procedure. They are illustrated in the following figures (p 101-102).

Figure 4 Illustration of the primary incision during the saphenofemoral junction ligation

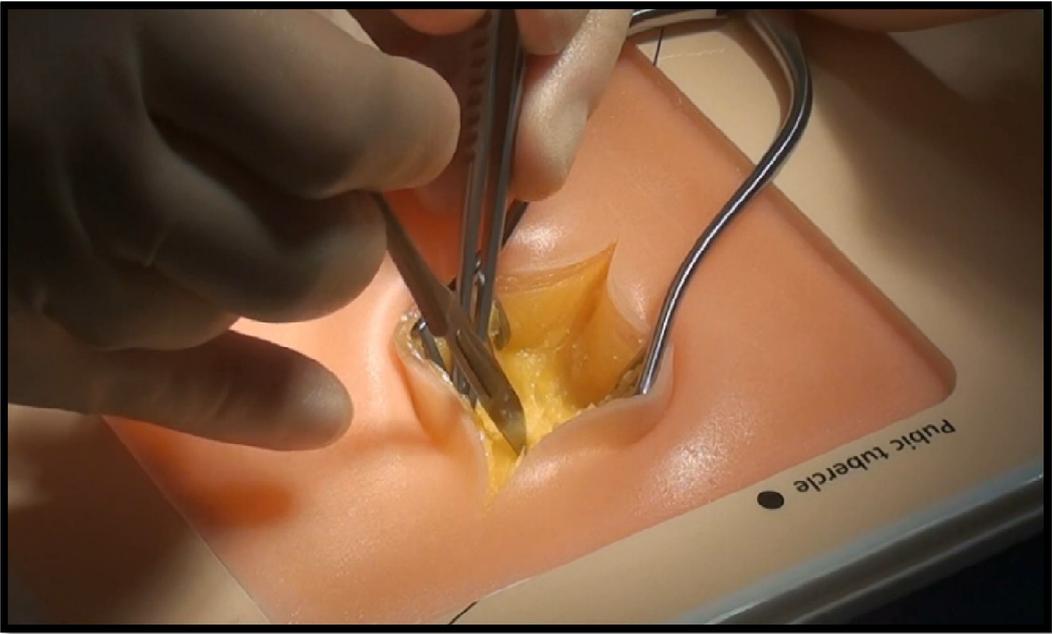
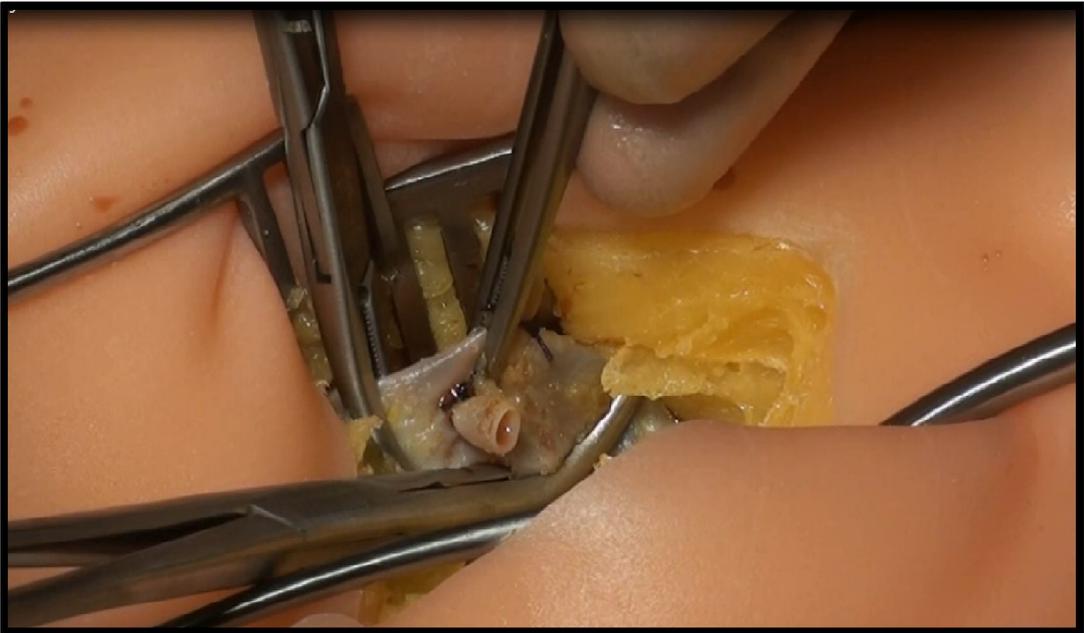


Figure 5. Illustration of ligated tributaries required during the dissection of the sapheno-femoral junction ligation

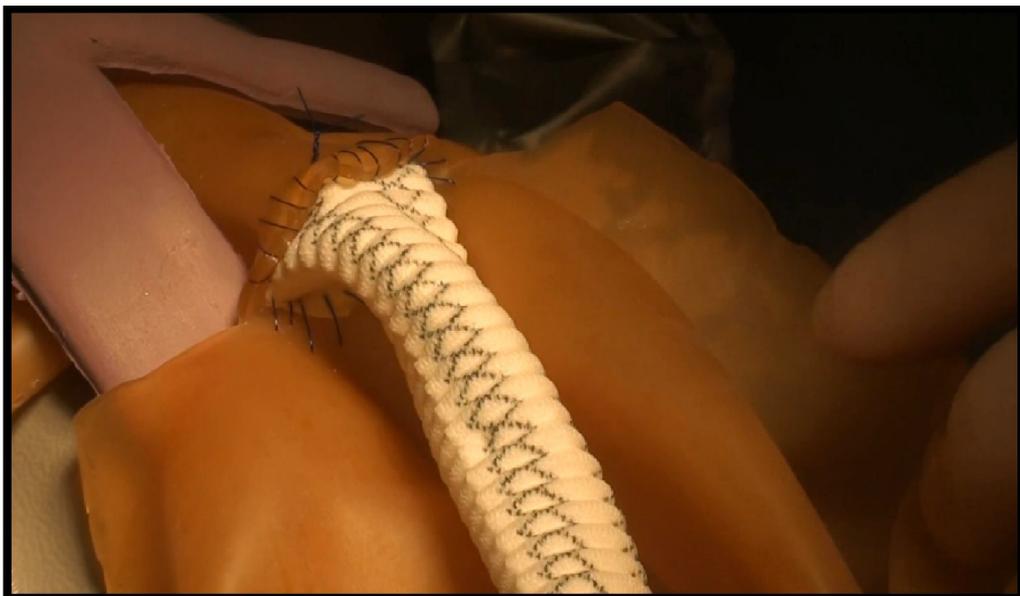


**Figure 6 Illustration of the aortic aneurysm simulator**

Inferior vena cava and left renal vein are illustrated abutting and crossing the proximal aortic respectively (pink vessels). The aneurysm sac has been opened and proximal aortic neck shown (orange vessel). The Dacron graft has been prepared and is about to be approximated to the neck of the aneurysm (white)



**Figure 7 Illustration of the completed proximal aortic anastomosis**



### **2.5.5 Study Protocol**

Many of the trainees who attended the course were of varying skill levels prior to performing the simulated procedures and they were all required to watch a standardised video illustrating the steps involved in the simulated operation before performing the procedure. The majority of the assessors and trainees had used PBA in the workplace they were already familiar with how the assessment tool was used and completed. An additional short briefing was conducted with study participants in the use of the second assessment tool, the generic section of the OSATS tool(101) prior to completing an assessment on a simulated procedure.

Finally each trainee performed a simulated procedure and was assessed by the trainee assisting them and by circulating assessors where possible. The forms were then gathered in hard copy by the study co-ordinator for analysis at a later date.

### **2.5.6 Good Clinical Practice and Ethical Considerations**

As the study included assessment of some cadaveric material, in addition to the data used and described in this experiment from the low fidelity simulator jugs, consultation was required from the Human Tissue Authority. The cadaveric data was used for a separate study. Permission from the licence holder within the dissection laboratories in the Royal College of Surgeons in London was required to conduct the study. This was to ensure the use of such materials in this fashion did not breach the restrictions of the Human Tissue Act 2004(181) and the study protocol underwent university ethical approval. The approval number from the Newcastle University Ethics Committee was 00530/2012.

Prior to collection of the data each trainee provided written consent to use their data in this study. Each trainee and assessor was approached individually by the study co-ordinator who had received current good clinical practice training at the start of the course. Course participants were briefed on what the study involved and the assessment tools used and to what use the data would be used. If a course participant declined involvement in the study they were not involved in the data collection or assessment process and reassured that non involvement would not adversely affect the outcome or completion of the course.

## **2.6 Experiment 2- Construct validity of the ANGIOMentor endovascular simulator and inter- rater reliability and responsiveness of Procedure Based Assessment in simulated superficial femoral artery angioplasty**

### **2.6.1 Study setting for experiment 2**

The setting for the second experiment within this study took place within the Hull and East Yorkshire NHS Trust and utilised the high fidelity endovascular simulator known as the ANGIOMentor. Based within the state of the art clinical skills facility it provided an ideal opportunity to assess the ability of the endovascular simulator to provide structured and assessed teaching sessions for interested trainers and trainees under study conditions. Again using an observational methodology trainees and trainers based within the trust were assessed performing a superficial femoral artery angioplasty procedure on the ANGIOMentor endovascular simulator.

### **2.6.2. Study participants**

Surgical trainees, vascular specialist nurses and vascular consultants were recruited to participate within this arm of the study. Seven novice trainees were initially recruited, who had no experience of arterial or venous interventional procedures. Five intermediate level trainees were also recruited which included 2 interventional radiology trainees whose experience included less than 100 superficial femoral artery angioplasties. The remaining 3 intermediate level trainees included 2 vascular trainees and 1 vascular consultant all of which had experience of greater than 200 endovascular venous ablation procedures. Finally 3 expert interventional

radiologists were recruited whose experience included greater than 500 superficial femoral artery angiograms and angioplasties.

### **2.6.3 Study Assessors**

In order to accurately assess trainees performing the interventional procedures, assessors of sufficient experience were required to provide an opinion of what would constitute an empirically correct performance on the endovascular simulator. Four vascular interventional radiologists were recruited to aid this process and provided assessments in conjunction with 3 additional experienced assessors. These included 2 vascular trainees with extensive experience on the endovascular simulator and 1 interventional radiology trainee with 2 years experience of radiology training and over 100 superficial femoral artery angiograms.

### **2.6.4 Materials and Procedures Assessed**

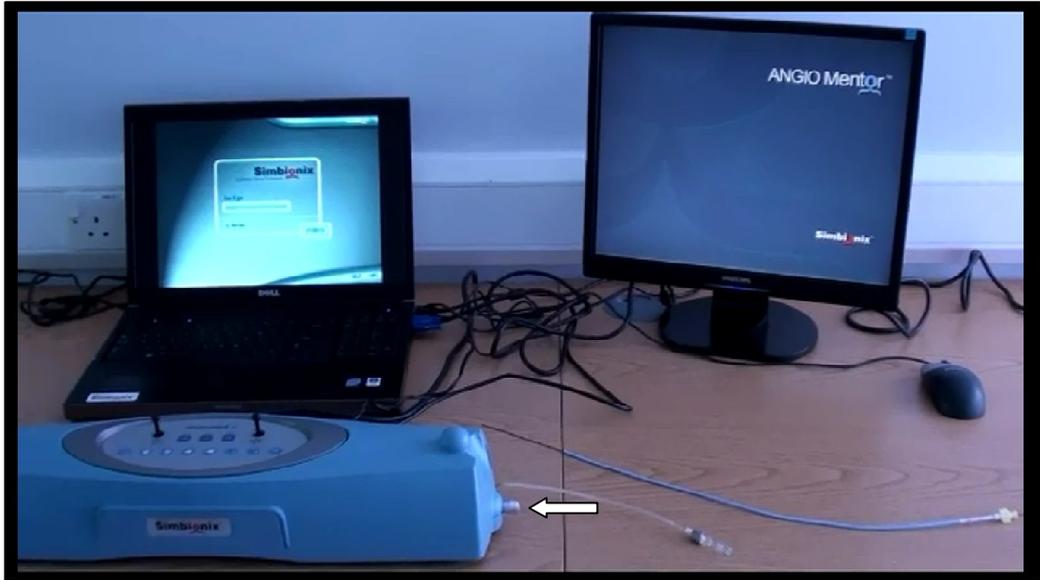
The procedure chosen to be the focus of this study and to determine if a modified PBA was a reliable and responsive measure of performance was a right superficial femoral artery angiogram and angioplasty. This is one of the first training modules that come pre-loaded on the software within the computer of the simulator. It was chosen as it represents one of the first procedures that an interventional trainee may be required to perform and utilises many of the skills required in interventional radiology such as wire manipulation and management, basic angiogram and angioplasty techniques.

The ANGIO Mentor™ simulator used was the first generation of simulators produced by Symbionix called the ANGIO Mentor™ Express. It is a portable unit which consists of the simulator unit and two visual monitor displays, one of which is generated by a laptop which contains the software and

learning modules for the unit. Once the simulator is in use it generates a fluoroscopy image on a larger screen and a visual read out of the drugs and patient vital signs on the other and an image of the patient on the x-ray table and x-ray tube so the user can monitor the position of these in relationship to the patient. (Figures 8-10 p108-109) X-rays are delivered by the operator using foot pedals so freeing the user to manipulate wires and catheter simultaneously.

**Figure 8 the ANGIOMentor simulator illustrating the two display readout and joystick controls for the x-ray tube and table**

**Illustrated also is the small metal docking port at the end of the blue simulator unit which represents a femoral access point (white arrow)**



**Figure 9 Illustration of the catheters and wires used to insert into the simulator to perform the interventional procedure of choice**

**Illustrated also are the syringes used for injecting contrast (small) and the angioplasty device (large syringe)**

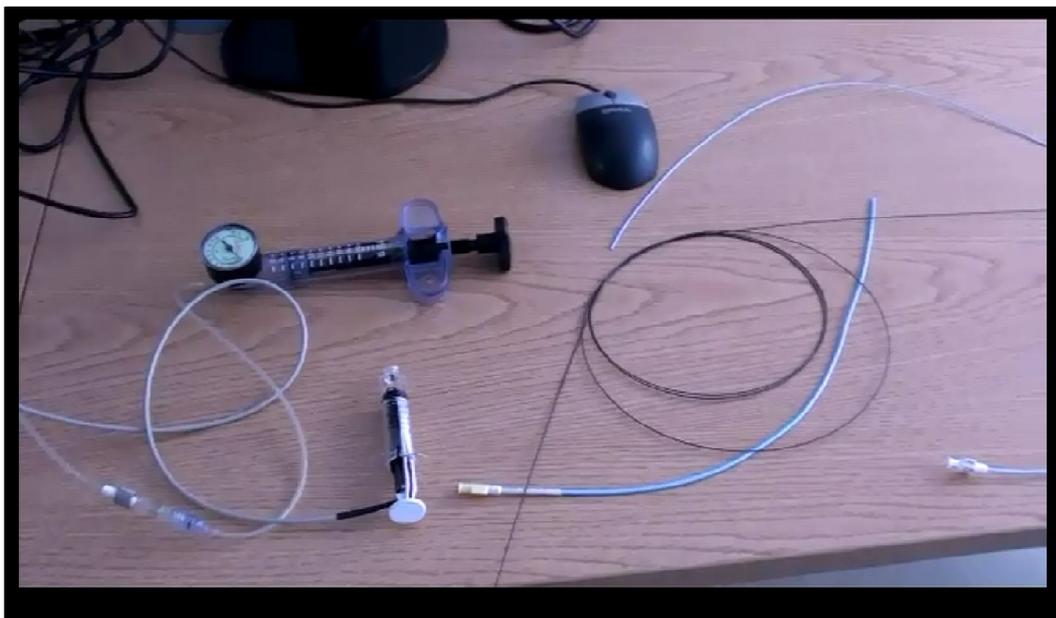
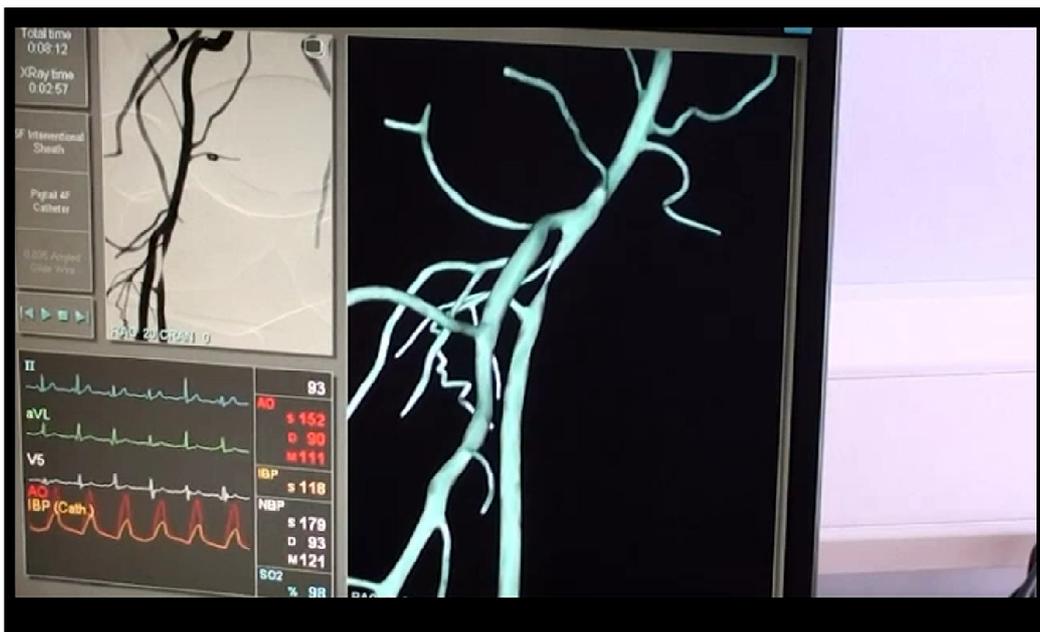


Figure 10 an example of the fluoroscopy readout after contrast injection of the left femoral access point

The catheter can just be seen in the bottom right hand corner in the left common femoral artery



Figure 11 illustrating the simulator's ability to rotate the lesion of interest in 3 D to aid teaching and catheter and wire positioning



Uniquely the unit is able to generate tactile feedback to the operator so that resistance can be felt as the user manipulates oversized instruments or through stenotic lesions. Figure 11 (p109) illustrates the lesion used in the study which the participants were required to identify and perform as part of the study.

### **2.6.5 Study protocol**

This experiment was split into 2 phases; the first phase was designed to determine if the ANGIOMentor could be a valid training tool and important component to ensuring the resulting data from the second phase was reliable. This phase of the study was included as this particular endovascular simulator and procedure had never been validated before. The second phase, which was the focus of this study, was to determine if a modified PBA was a reliable and responsive assessment tool in assessing trainees performing on the simulator.

Prior to performing on the simulator each participant was given a short didactic lecture on how the simulator instruments were used and how contrast and x-ray screening could be delivered on the simulator. They also received a demonstration of the joystick controls to allow for movement of the patient and x-ray tube as required. They also received a copy of a step by step guide produced by a consultant interventional radiologist on how to perform the procedure. This was available to the participants throughout the procedure to refer to if they wished and was designed to provide trainees with an example of empirically correct performance designed to highlight safe practice and avoid complications.

In order to demonstrate the ANGIOMentor as a valid training tool, 3 groups of study participants were recruited. Novices, intermediate level participants and experts performed 5 right superficial femoral artery

angiograms and angioplasty of a TASC A lesion situated just below the common femoral artery. The data metrics then generated by the simulator were then used to determine if there were demonstrable differences between groups, which would confirm the simulator as a valid training tool. The second phase and primary outcome of interest, of the study utilised the data and performances of the novice group of trainees. In this phase of the study, the 5<sup>th</sup> run performed by the trainees was subject to an observational assessment. The novice trainees were then reassessed within 3 months where they performed an additional 5 runs and were again assessed on their 10<sup>th</sup> and final run on the simulator. This allowed us to demonstrate if the simulator could be a responsive tool in assessing performance over time.

#### **2.6.6 Good clinical practice and ethical considerations**

Prior to undertaking experiment 2 formal ethical approval was required by the Hull and East Yorkshire NHS trust research and development department as the study involved employees of the trust. Therefore a study protocol and documents were submitted for proportionate review to Derby ethics committee and approval given for the study to proceed with the study, research ethics committee number 11/EM/0317. All participants including the assessors involved in the study gave formal written approval for study inclusion.

## **2.7 Experiment 3- Intra-rater Reliability of Procedure Based Assessment in simulated vascular procedures.**

### **2.7.1. Study Setting for experiment 3**

The final experiment within this study was conducted to explore whether the PBA tool could demonstrate intra-rater reliability, the property of a tool to determine if there agreement by a single rater over multiple repetitions of the same test(45). The experiment was conducted within the vascular department of Hull Royal Infirmary in the Hull and East Yorkshire NHS trust.

### **2.7.2 Study participants**

This study utilised a single videotaped performance of an experienced vascular surgeon performing a simulated sapheno-femoral junction dissection and ligation with the aid of an assistant, both participants were anonymised for the purpose of the study and wore surgical gloves and gowns to help protect their identities. (The video is submitted as an appendix)

### **2.7.3 Study assessors**

The assessors involved in the study included vascular consultants a, surgical and vascular trainees and experienced vascular nurses who had regularly assisted with sapheno-femoral junction ligation and varicose vein surgery.

### **2.7.4 Materials and procedures assessed**

A single videotaped performance was recorded in the minor operations and day procedures unit in an empty operating theatre using a high

definition Sony Handycam™ with a 25 times optical zoom to produce a high quality video. The simulator used was the same as described in experiment 1, the Limbs and Things™ single use low fidelity synthetic sapheno-femoral junction ligation jig. All instruments and sutures used were those that would be used in a standard sapheno-femoral junction and instrument tray. Once the procedure was completed the video was edited by the study coordinator using commercially available Sony Vegas Movie HD Platinum version 10.0 video editing package to a 16 minute 30 second video which included the entire procedure from exposure to closure. All sound was removed from the video to help protect the anonymity of the operators.

### **2.7.5 Study protocol**

Once the video had been edited it was loaded on CD to enable playback at a convenient time. Two initial sessions were set, 6 weeks apart, in the regular departmental Monday morning meetings to allow for playback and the initial assessment session. This was a pragmatic decision which allowed for the maximum number of raters to be available at any one time. This was then followed by the second session 6 weeks later where the video was played a second time and the raters performed a second assessment of the video. Periods between to 2 weeks (182) to 6 months(183) in health and psychological measurement have shown good test –retest reliability. Therefore 6 weeks was deemed to be a reasonable time frame within the period of the study to ensure collection of the second paired data set prior to rotation of junior raters from one post to the next. Where raters were available from the first sitting, but not for the second, they were individually contacted and supplied with a copy of the video and assessment tools so they could perform the second assessment in a timely manner at their convenience.

### **2.7.6 Good clinical Practice and ethical considerations**

Written permission was given by the surgeon illustrated in the video for its use within this experiment and all efforts were made to anonymise the surgeon involved. No patients or other study subjects were involved as the focus of assessment. All assessors gave verbal permission to use their assessments within this study and were only identified by training level.

## **Chapter 3- Results**

### **3.1 Experiment 1- Validity and Reliability of Procedure based assessment in simulated vascular procedures**

#### **3.1.1 Participant recruitment and baseline characteristics**

In total there were 22 individual trainees who agreed to be assessed for inclusion in the study. This represented 22 individual procedures which were dually assessed in 3 cases. Therefore in total there were twenty five individual assessments performed of which 11 assessments took place using the simulated sapheno-femoral junction jig (Limbs and Things) and 14 assessments which utilised the aortic aneurysm proximal anastomosis jig (Limbs and Things). All participants were male and training grades ranged from ST1 equivalent to ST5 equivalent. Assessors also ranged from ST1 to ST5 but included a significant number of consultant assessors. The distribution of the grade assessors and participants' is illustrated in table 9 (page 117) The data was then analysed for normality and determined to be non gaussain in distribution therefore non-parametric tests were used to analyse the data. Validity was determined by calculating the spearman correlation co-efficient for the scores from the modified PBA global and task specific checklist element of PBA (GTSC) and the global summary score (GSS) in relationship to:

- Number of procedures previously observed and performed

- The level of training grade

- In addition the results of the above variables will be presented from the OSTATS assessment tool to demonstrate if PBA has similar levels of validity and whether the scores between these instruments correlate to a

significant degree suggesting concurrent validity. Where data is illustrated graphically trend lines have been added where applicable based on the data. Where data is expected to model a learning curve, trend lines are non-linear.

**Table 9 Distribution of assessors and trainees performing sapheno-femoral junction ligation and proximal aortic anastomosis**

Median PBA and OSATS scores and previous operative experience are also given for the group and individual trainee level.

	Assessor designation N=25	Procedures Assessed N=25	Participants N=22	Median PBA adjusted GTSC score for all assessed procedures IQR=() N=25	Median PBA GSS score for all assessed procedure IQR=() N=25	Median OPS score for all assessed procedures IQR () N=23	Median Number of procedures previously observed for all assessed procedures IQR () N=24	Median number of procedures previously performed for all assessed procedures IQR= () N=24
ST1	3	6	5	0.650 (0.545-0.758)	2 ( 1.5-2.25)	22 (17.50-27.25)	10 (1.5 – 25)	0 (0-0.50)
ST2	1	3	2	0.720 (0.660-0.792)	3 (2-3)	21 (19.00-21.00)	27(10-27)	10 (0-10)
ST3	1	4	3	0.765 (0.600- 0.952)	2.5 (.50- 3.75)	21 (13.00-21.00)	25 (12.50-30)	3 (0-6.0)
ST4	10	9	9	0.960 (0.795- 1.00)	3.0(2.50-3.5)	33.00 (30.50-35.00)	30 (11-65)	2 (0.5-35)
ST5	2	3	3	0.840 (0.650-0.968)	2.0 (2-3.60)	28.00 (25.00-28.00)	25 (1.0- 29)	4 (1-4)
Cons	8	N/A	N/A	N/a	N/A	N/A	N/A	N/A
Group Medians and IQR ()	ST4 (ST4-Consultant)	ST3 (ST2-ST4)	ST4 (ST2-ST4)	0.790 (0.675-0.990)	2.5 (2-3)	27.00 (21.00-32.00)	25 (10-30)	1.5 (0-9)

### **3.1.2 Validity of Global and Task Specific Checklist (GTSC) and Global Summary Score (GSS) of Procedure Based Assessment for number of procedures previously observed**

The correlation co-efficients for the GTSC and GSS scores are based on data from 24 individual assessments. Unfortunately one trainee, while supplying their training grade did not supply their previous operative experience therefore it was omitted from the data analysis to determine the validity of the GTSC and GSS. The correlation co-efficient of the GTSC for number of operation previously observed was  $r = 0.294$  ( $p = 0.163$ ). The co-efficient for the GSS for number of operations previously observed was  $r = 0.335$  ( $p = 0.110$ ). The results are illustrated on the following page. Despite illustrating a positive correlation with increasing observational exposure the main finding was a non-significant relationship between observation of procedures for the GTSC and GS for the PBA tool.



### **3.1.3 Validity of Global and Task Specific Checklist (GTSC) and Global summary score (GSS) of Procedure Based Assessment for number of procedures previously performed.**

The Spearman's Rank Correlation co-efficient for the PBA GTSC for number of operations previously performed was  $r= 0.446$  ( $p=0.029$ ). The correlation co-efficient for the PBA GSS for operation previously performed was  $r=0.553$  ( $p= 0.005$ ). The results are described graphically on the following page. In contrast to observation when the GTSC and GSS are correlated with numbers of operations previously performed there is a strong statistically significant trend suggesting construct validity for PBA for this variable.

Figure 14 Scatter graph illustrating the distribution of PBA GTSC scores and Spearman's correlation co-efficient with number of operations previously performed

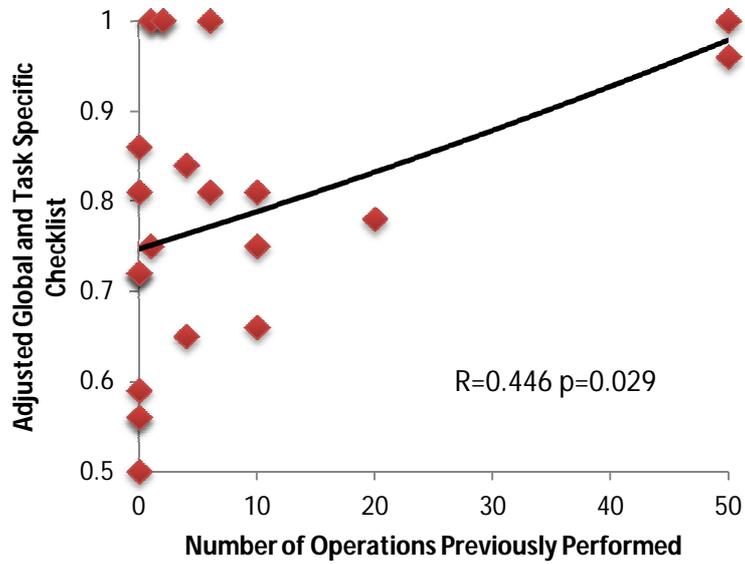
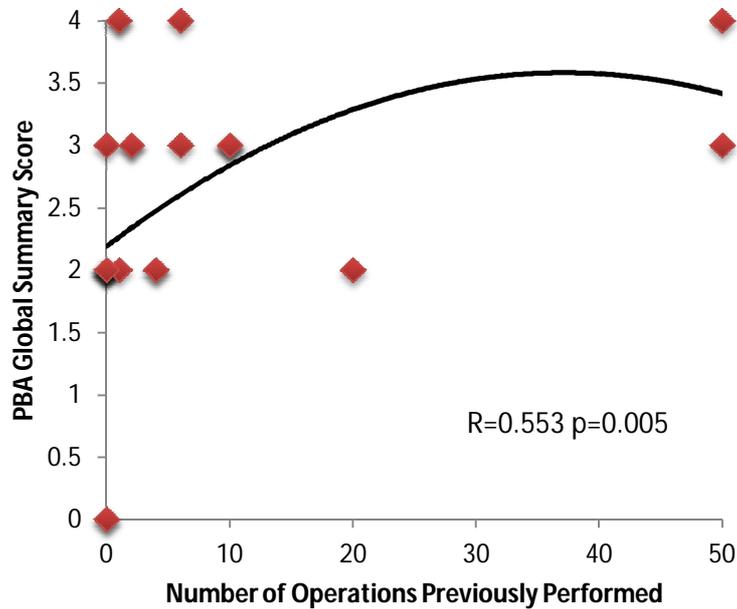


Figure 15 Scatter graph illustrating the distribution of PBA GSS and Spearman's correlation co-efficient with number of operations previously performed



### **3.1.4 Validity of the Global and Specific Checklist and Global Summary Score of Procedure Based Assessment in relationship to training grade**

The results from the SFJ and Aortic Jig performances were correlated with the training grade of those trainees who performed these simulated procedures. This demonstrated a moderate and significant correlation with the GTSC ( $r= 0.588$   $p=0.002$ ) and a lower and borderline significant correlation the GSS ( $r=0.352$ ,  $p=0.088$ ). The results are illustrated graphically on the following page.

Figure 16 Scatter graph illustrating the distribution of PBA GTSC and Spearman's correlation co-efficient with year in surgical training

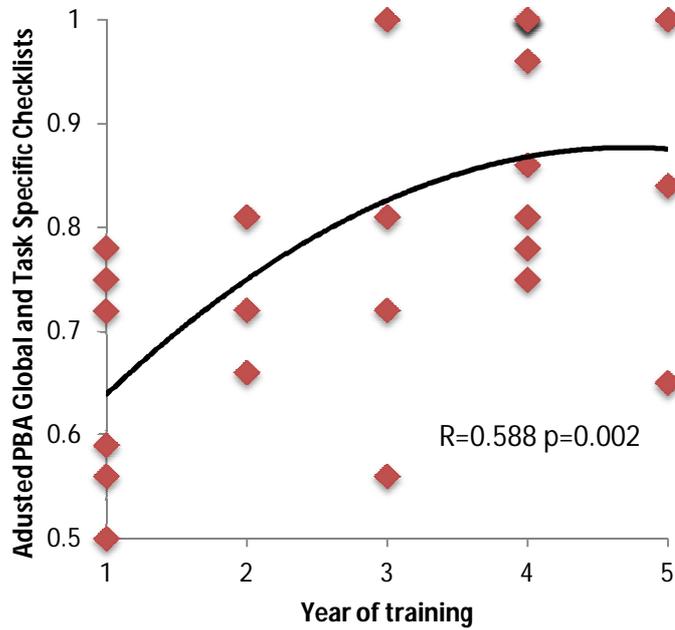
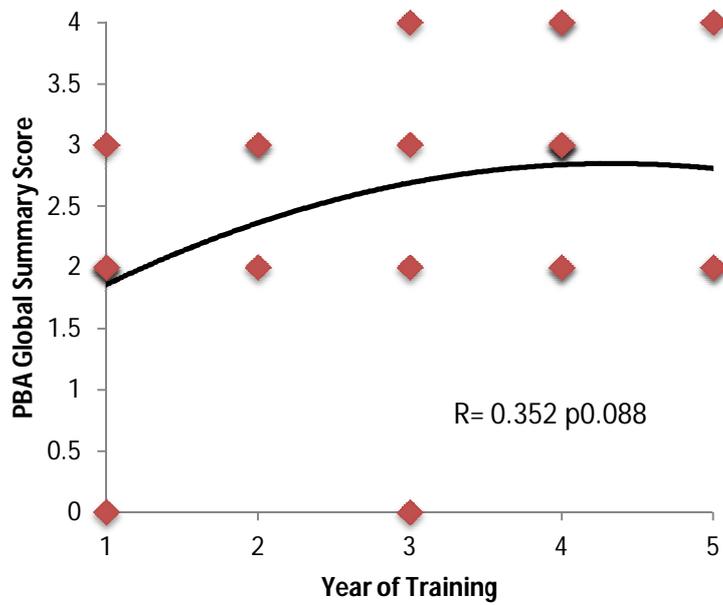


Figure 17 Scatter graph illustrating the distribution of PBA GSS and Spearman's correlation co-efficient with year of surgical training



### 3.1.5 Validity of Objective Structured Assessment of Technical Skills in assessment of previous observation, performance and training grade

In order to compare PBA and determine if this newer assessment tool could have similar or improved validity as the previously validated OSATS assessment tool the variables of observation, performance and training grade were analysed. These variables were analysed with respect to the OSATS score given by the same assessors for the above procedures. When the OSATS scores were correlated with number of operations previously observed ( $r=0.231$ ,  $p=0.302$ ) and performed ( $r=0.347$ ,  $p=0.114$ ) the correlation values were low and non significant. However when the OSATS assessment results were correlated with training grade the result demonstrated a moderate correlation that was statistically significant ( $r=0.634$ ,  $p=0.001$ ). The results are illustrated graphically on the following pages.

Figure 18 Scatter graph illustrating the distribution of OSATS scores and Spearman's correlation co-efficient with number of operations previously observed

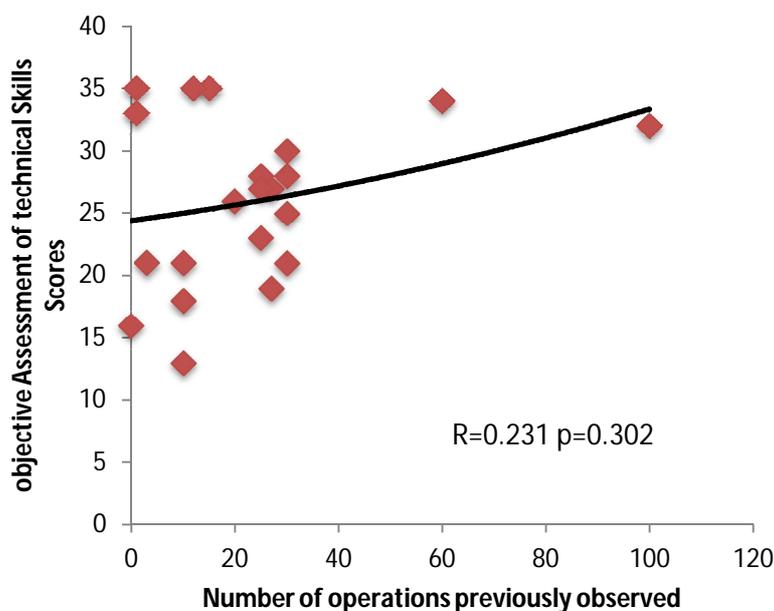


Figure 19 Scatter graph illustrating the distribution of OSATS scores and Spearman's correlation co-efficient with number of operations previously performed

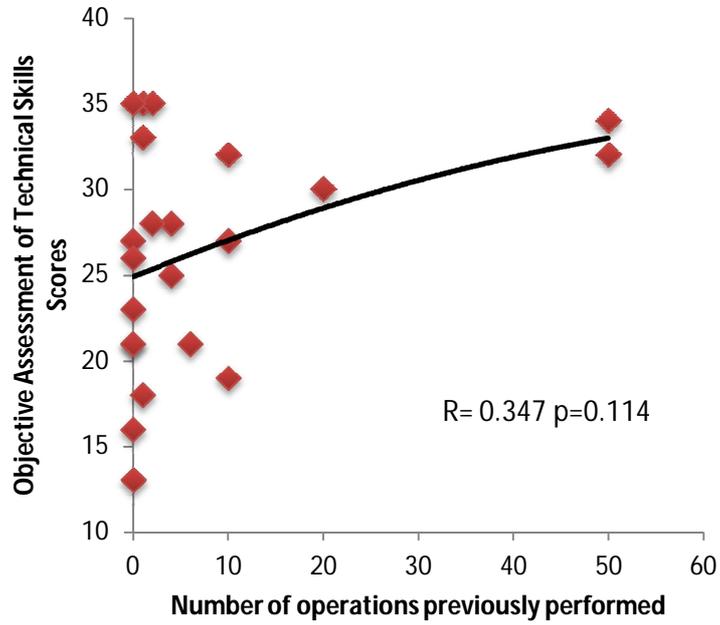
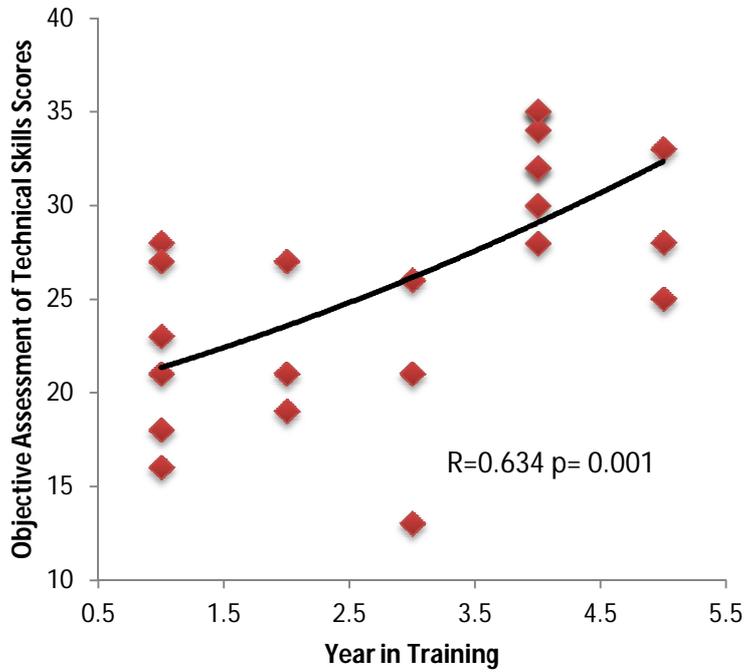


Figure 20 Scatter graph illustrating the distribution of OSATS scores and correlation co-efficient with year in surgical training



### 3.1.6 Concurrent validity between Objective Assessment of Technical Skills and the Procedure Based Assessment

An important aspect of validity is to determine if an assessment tool possesses concurrent validity with a previously validated tool. Bland and Altman recommend that to determine whether two different types of measurements agree especially when they use different methods, a plot of the difference between the methods against their means can help to show discrepancies. Simple correlation will only determine if the different measures are linearly related not if the two measures agree(180). The results are illustrated graphically on the following pages. The one sample t-test has been calculated to determine the null hypothesis that there will be zero difference between the means.

Figure 21 Bland-Altman Plot describing the relationship and level of agreement between OSATS and the ATIS PBA Scores.

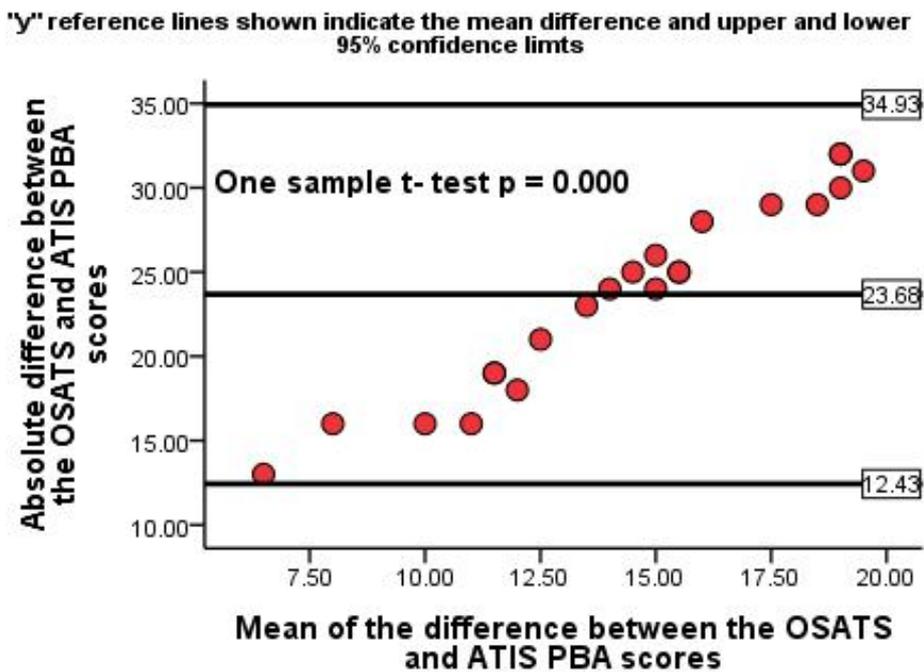
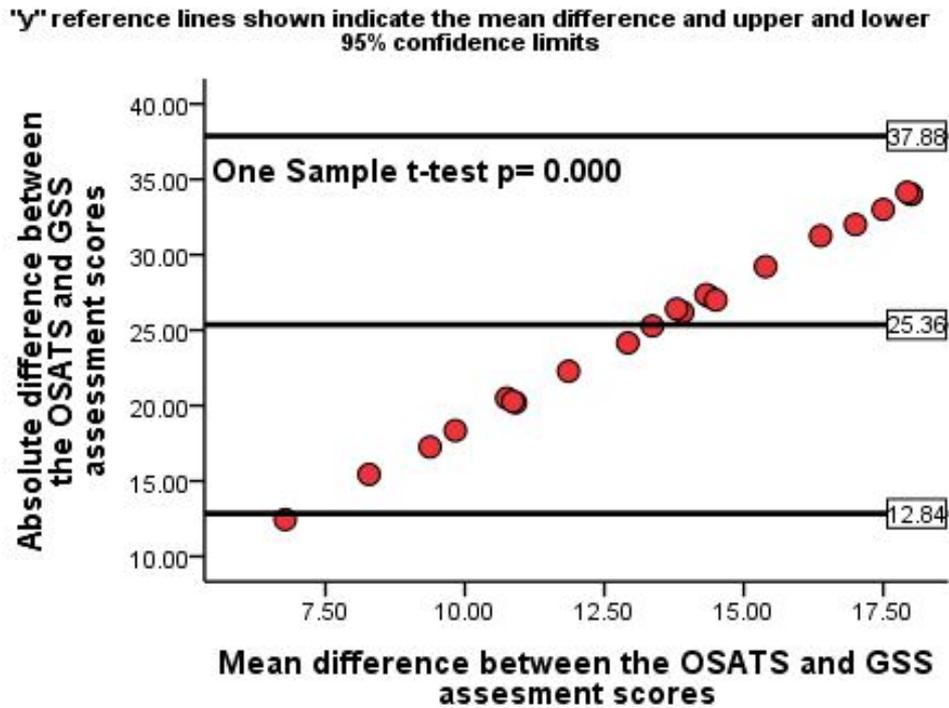


Figure 22 Bland-Altman Plot illustrating the relationship and level of agreement between OSATS and the PBA GSS



Here the one sample t-test has shown that there is a significant difference between the two measurements; hence poor agreement. The Bland-Altman plots illustrate that there are bigger differences between the measurements the larger the mean of the two, indicating there are higher levels of error between larger scores. There are also wide confidence limits indicating poor agreement.

## **3.2 Experiment 2- Inter-rater Reliability and Responsiveness of Procedure Based assessment in simulated vascular procedures**

### **3.2.1 Part a) Determining the validity of the ANGIOMentor Simulator**

While many of the bench top simulators such as the SFJ jig and aortic jig had previously been validated as a training tool in previous studies the ANGIOMentor is a relatively new simulator and had never been validated before in superficial femoral artery angioplasty. Therefore the validity of the simulator to distinguish between different levels of trainee was established using the data metrics produced by the simulator. It must be noted that data metrics provided by such simulators may only translate in a limited way to a real world scenario. Such metrics as time to complete procedure are only surrogate markers of performance and can only be indirectly linked to procedural flow and skill. As a training tool this may not be directly useful to the novice operator, time is much less important than undertaking the procedure in a safe and error free manner. Metrics that can distinguish and highlight errors are potentially much more useful in a training context. As outlined in chapter 1, metrics that highlight empirically correct performance such as handling errors and quality of final product are potentially the most useful. In this experiment the ANGIOMentor was found to be valid between novices, intermediates and experts in total fluoroscopy time, total procedure time and time to position sheath for intervention ( $p < 0.001$ ). Residual stenosis was significantly different between novice and intermediates (MWU  $p = .012$ ) but not between intermediates and experts (MWU  $p = .303$ ). Contrast use was only significant between novice and intermediates (MWU  $p = .020$ ) and intermediates and

experts (MWU  $p = .018$ ). There were no significant differences between groups in advancement of the catheter or sheath without a leading wire. The main findings are illustrated in figures 23-28 (pages 130-132).

Of those metrics that are provided by the ANGIOMentor simulator clearly metrics such as contrast use and fluoroscopy use can provide useful training feedback and have been shown to be valid or partially valid in the case of contrast use within this study. The ideal simulator metric would of course be able to give the trainee feedback on the quality of final product and highlight wire handling errors; which are only partially covered here in examining travel of the leading wire without a sheath and residual stenosis

Box plots illustrating and comparing the performance of Novice (red), Intermediate (blue) and Expert (green) participants across a range of simulator generated metrics. Bars indicate median values, boxes interquartile range and whiskers the statistical range. Circles represent outliers. Kruskal-Wallis statistical test used unless indicated.

Figure 23 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for total procedure time

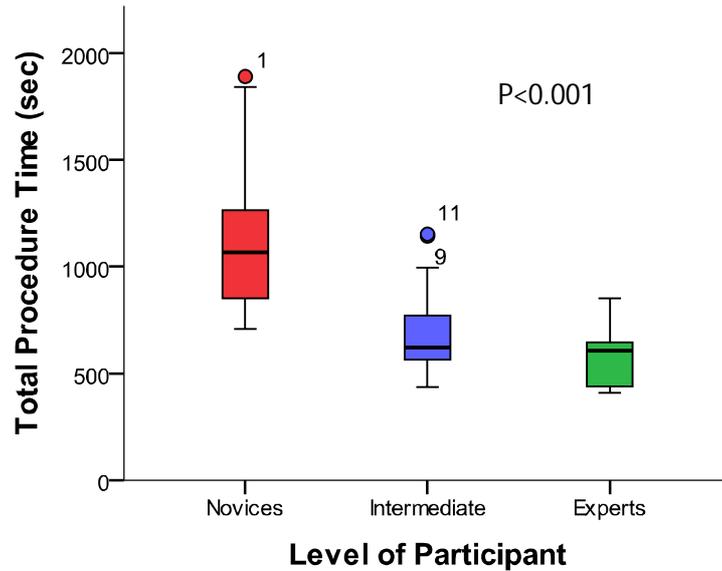


Figure 24 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for total fluoroscopy time

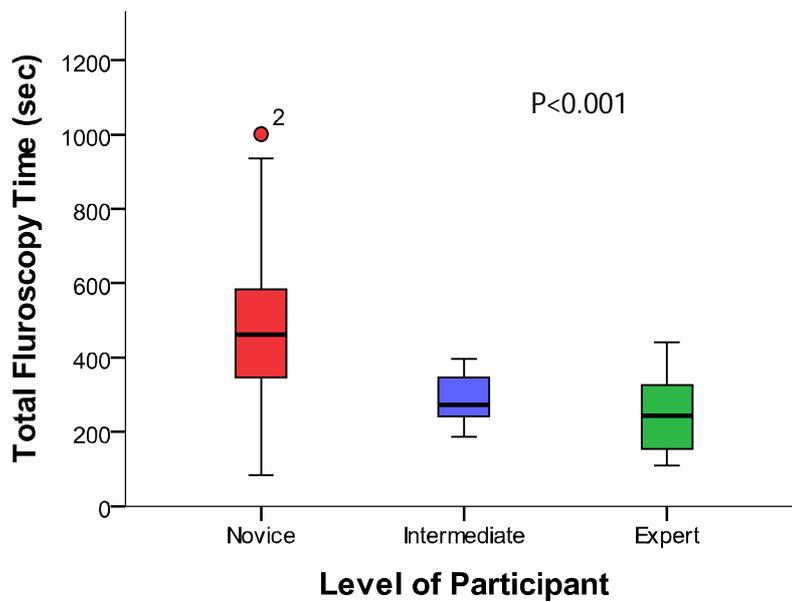


Figure 25 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for time to position sheath for intervention

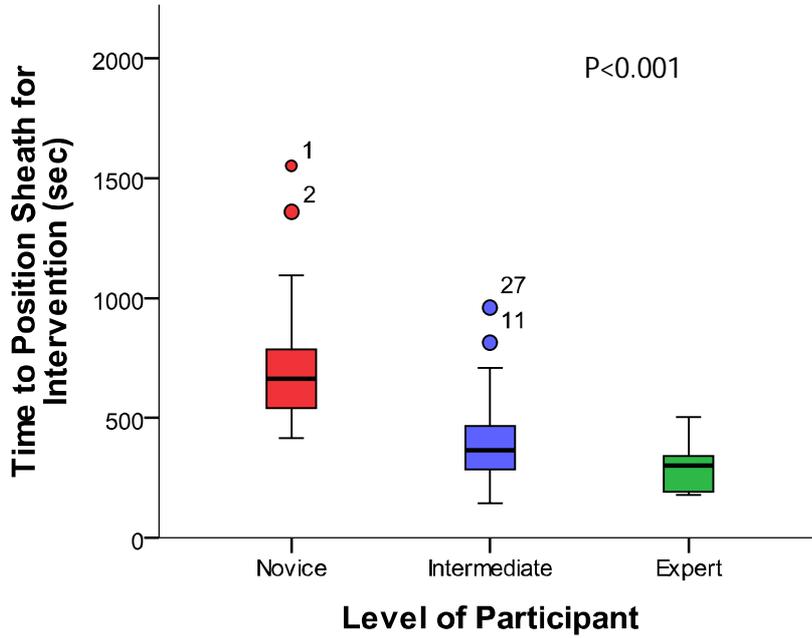


Figure 26 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for residual stenosis of the SFA lesion after angioplasty

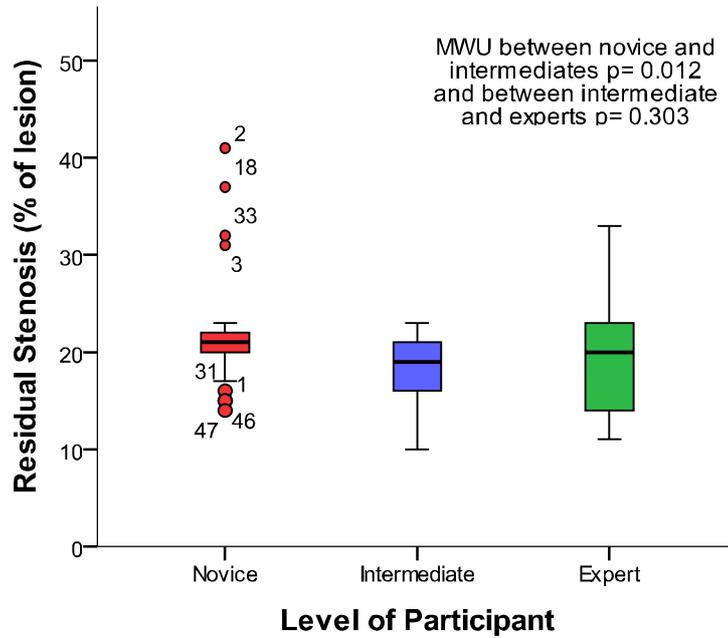


Figure 27 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for contrast volume used

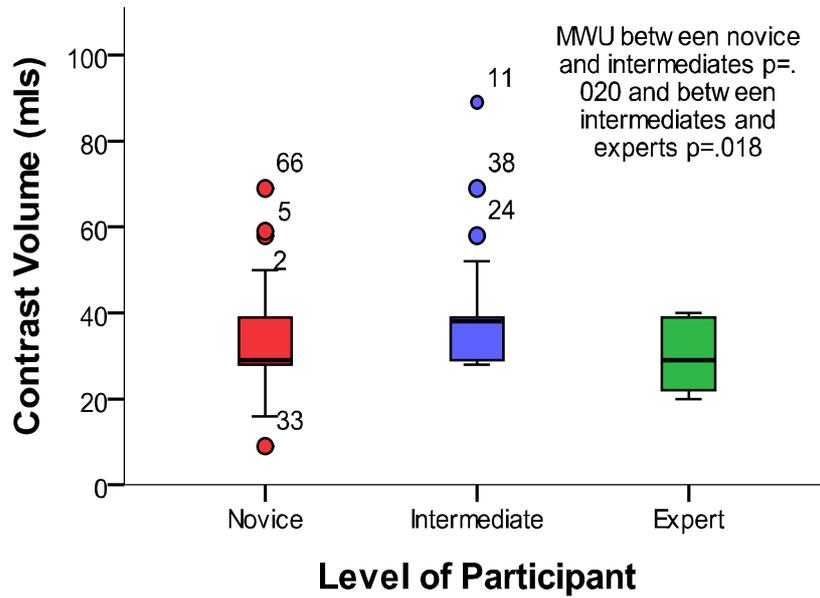
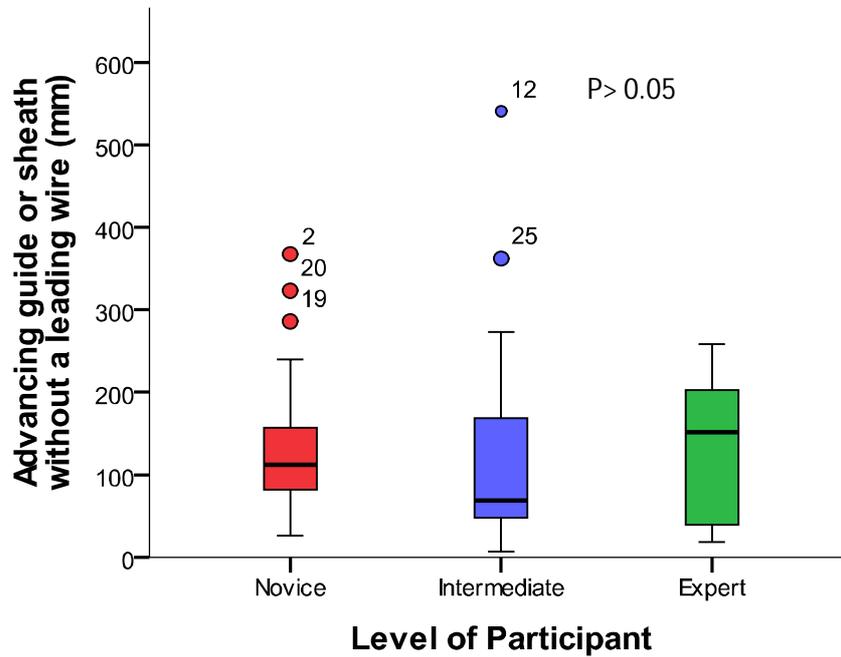


Figure 28 Box and whisker plot illustrating participant performance on the ANGIOMentor simulator for advancement of the guide or sheath without a leading wire



### **3.2.2 Part B) Participant recruitment and baseline data for Inter-rater reliability and responsiveness**

The experiment to demonstrate if PBA possessed inter-rater reliability; the agreement of two raters assessing the same construct simultaneously, was originally designed to be derived from data in experiment 1. However due to difficulties with obtaining multiple raters for each of the 22 individual procedures the reliability data from experiment 1, (2 sapheno-femoral junction jig procedures and 1 aortic jig procedure) was combined with the data from experiment 2. Seven novice trainees agreed to participate in experiment 2. They included 2 vascular research nurses, 2 CT2 surgical trainees and 3, CT2 equivalent research fellows. There were 3 female and 4 male participants. The data from experiment 2 provided 13 individual novice procedures from 7 individuals, 6 of which performed an assessed run twice. The ANGIOMentor™ assessments were in the majority of cases performed by two individuals; in some cases 3 assessors were present for each of the assessed simulations. This allowed 13 pairs of assessments from experiment 2 to be combined with 3 pairs of assessments from experiment 1. Despite the different nature of the procedures; it is the actual assessment method under scrutiny rather than the procedure itself, the GTSC checklist was analysed in the same way to give a ratio of the perfect score despite differences in the number and definition of items within the checklist so allowed parity between experiments.

The raw data scores for all the procedures utilised to assess inter-rater reliability are given in the table 10 (p 135).

As previously the distribution of scores from both elements of the PBA assessment tool were analysed for normality and shown to be non-gaussian in distribution therefore Spearman's Rank correlation co-efficient was utilise to calculate the reliability of both elements of the PBA

assessment tool and Cronbachs Alpha was determined to give an estimate of hypothesis free reliability.

**Table 10 Original data from all paired PBA assessments from Experiment 1 and Experiment 2**

Participant	Training level	Assessment 1		Assessment 2	
		Adjusted GTSC	GSS	Adjusted GTSC	GSS
Novice 1	ST2	0.85	2	1.00	2
Novice 2	ST2	0.39	0	0.53	1
Novice 2	ST2	0.53	1	0.61	1
Novice 2	ST2	0.61	1	0.39	0
Novice 3	ST2	0.46	1	0.64	1
Novice 3	ST2	0.82	2	0.85	2
Novice 4	ST1	0.53	2	0.46	2
Novice 4	ST1	0.57	1	0.5	2
Novice 5	Vascular Nurse	0.39	2	0.43	2
Novice 6	Vascular Nurse	0.64	1	0.64	1
Novice 6	Vascular Nurse	0.42	2	0.71	2
Novice 7	ST1	0.61	1	0.64	2
Novice 7	ST1	0.71	2	0.75	2
SFJ 1	ST2	0.81	3	0.66	3
SFJ2	ST3	0.81	3	1.00	4
AAA 1	ST1	0.59	2	0.72	2

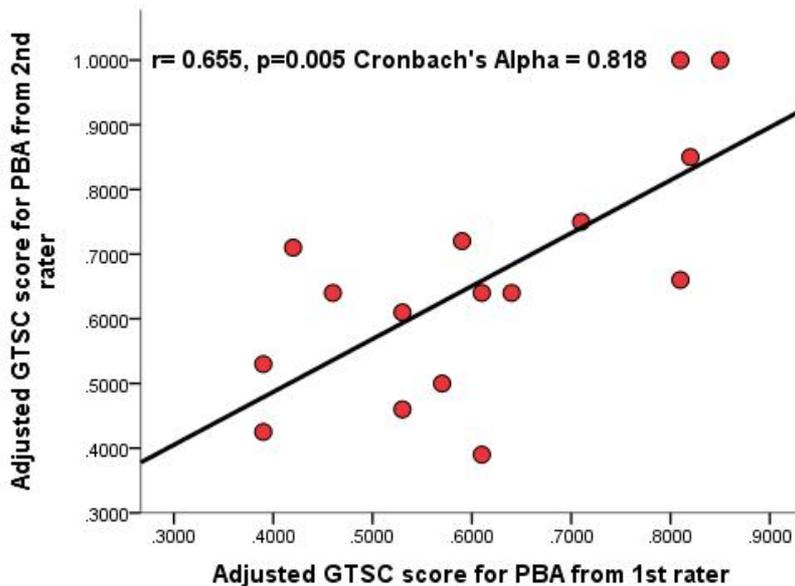
Each paired assessment has been generated by utilising all assessments performed for each trainee. Where there have been more than two assessors for each single procedure as is the case with Novice 2, all three assessments have been compared to one another to estimate reliability.

### 3.3 Inter-rater reliability

#### 3.3.1 Inter-rater reliability of the Procedure based assessment global and task specific checklist (GTSC) for simulated vascular procedures

The Spearman's rank correlation co-efficient for the GTSC element of the PBA tool was  $r = 0.665$  ( $p = 0.005$ ), indicating a moderate and statistically significant level of reliability. Cronbach's alpha was determined to be 0.818 suggested a high level of inter-rater reliability between the two assessors with minimal elements of variance. The data is graphically illustrated below.

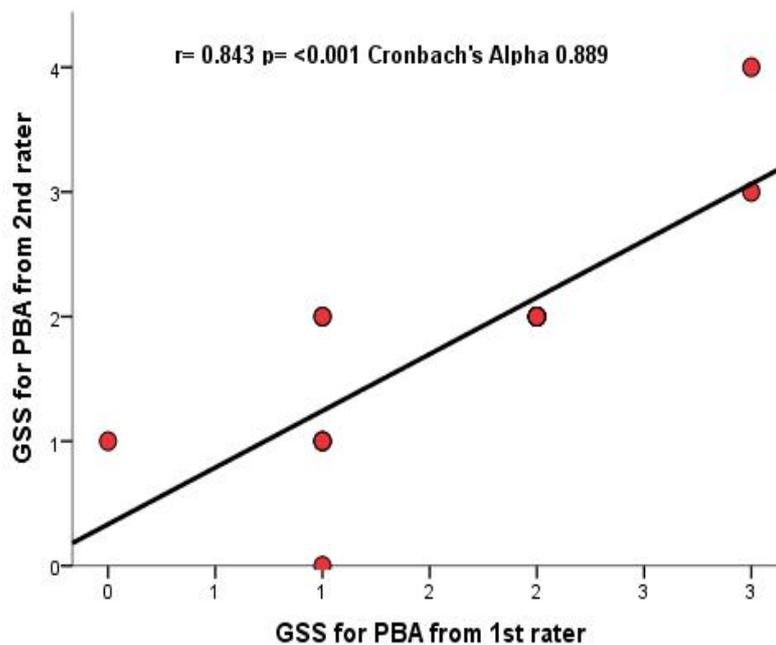
Figure 29 Scatter graph illustrating the distribution of PBA GTSC scores and Spearman's correlation co-efficient between 1<sup>st</sup> and 2<sup>nd</sup> raters



### 3.3.2 Inter-rater reliability of the Procedure Based Assessment Global Summary Score (GSS) simulated vascular procedures

The Spearman's rank correlation co-efficient for the GSS element of PBA tool was high and reached statistical significance  $r= 0.843$   $p> 0.001$ . Cronbachs alpha coefficient was estimated at 0.889 confirming a high level of inter –rater reliability between the two assessors. The data is graphically illustrated below.

Figure 30 Scatter graph illustrating the distribution of PBA GSS and correlation co-efficient between 1<sup>st</sup> and 2<sup>nd</sup> raters



### **3.4 Responsiveness of Procedure Based Assessment in simulated vascular procedures**

Only 6 out of the 7 novice trainees recruited completed the entire study protocol which consisted of 10 simulated runs on the ANGIOMentor endovascular simulator. The participant who failed to complete the second simulator session and second assessment (novice 2) was excluded from the responsiveness analysis. The distribution of assessments and raw scores are illustrated in table 11 (p 139). Each primary assessment was compared to each secondary assessment where applicable to give 22 pair wise comparisons. The data is illustrated in table 12 (p 140) and graphically illustrated in figures 31-32 (p 141). The main finding was that the adjusted global and task specific checklist was not a responsive element of the PBA tool ( $p=0.104$ ), However the PBA global summary score was responsive to increasing experience ( $p <.001$ ).

Table 11 Primary data from experiment 2 illustrating individual participant performance between the 5<sup>th</sup> and 10<sup>th</sup> simulator run on the ANGIOMentor simulator

Participant designation	1 <sup>st</sup> Assessed run (5 <sup>th</sup> simulator run)						2 <sup>nd</sup> Assessed run (10 <sup>th</sup> simulator run)				Number of days between assessments  (median 94 days (IQR 40-112 days))
	1 <sup>st</sup> rater		2 <sup>nd</sup> rater		3 <sup>rd</sup> rater		1 <sup>st</sup> rater		2 <sup>nd</sup> rater		
	Adjusted GTSC	GSS	Adjusted GTSC	GSS	Adjusted GTSC	GSS	Adjusted GTSC	GSS	Adjusted GTSC	GSS	
<b>Novice 1</b>	0.64	2	No rater available				0.85	2	1.00	2	112
<b>Novice 2</b>	0.39	0	0.53	1	0.61	1	Study drop out				
<b>Novice 3</b>	0.46	1	0.64	1			0.82	2	0.85	2	114
<b>Novice 4</b>	0.57	1	0.5	2			0.53	2	0.46	2	84
<b>Novice 5</b>	0.67	1	No rater available				0.39	2	0.425	2	42
<b>Novice 6</b>	0.64	1	0.64	1			0.42	2	0.71	2	37
<b>Novice 7</b>	0.61	1	0.64	2	0.46	1	0.71	2	0.75	2	104

**Table 12 all matched pairs of PBA on the ANGIOMentor (1st and 2nd assessed runs)**

Participant	1 <sup>st</sup> Assessment		2 <sup>nd</sup> Assessment		Wilcoxon Rank Sum Test (p values)	
	Adjusted GTSC	GSS	Adjusted GTSC	GSS	Adjusted GTSC	GSS
Novice 1	0.64	2	0.85	2	P=0.180	P=1.000
	0.64	2	1.00	2		
Novice 3	0.46	1	0.82	2	P=0.068	P=0.046*
	0.46	1	0.85	2		
	0.64	1	0.82	2		
	0.64	1	0.85	2		
Novice 4	0.57	1	0.53	2	P=0.141	P=0.157
	0.57	1	0.46	2		
	0.50	2	0.53	2		
	0.50	2	0.46	2		
Novice 5	0.67	1	0.39	2	P=0.180	P=0.157
	0.67	1	0.43	2		
Novice 6	0.64	1	0.42	2	P=0.458	P=0.046*
	0.64	1	0.71	2		
	0.64	1	0.42	2		
	0.64	1	0.71	2		
Novice7	0.61	1	0.71	2	P=0.028*	P=0.046*
	0.61	1	0.75	2		
	0.64	2	0.71	2		
	0.64	2	0.75	2		
	0.46	1	0.71	2		
	0.46	1	0.75	2		
<b>Cohort medians</b>	0.64 (IQR 0.50-0.64)	1 (IQR 1-2)	0.71 (IQR-0.46-0.82)	2	P=0.104	P<0.001*

### 3.4.1 Responsiveness - Cohort data

Figure 31 Box and whisker graph illustrating the cohort scores for the PBA GTSC between the 5<sup>th</sup> and 10<sup>th</sup> run on the ANGIOMentor simulator

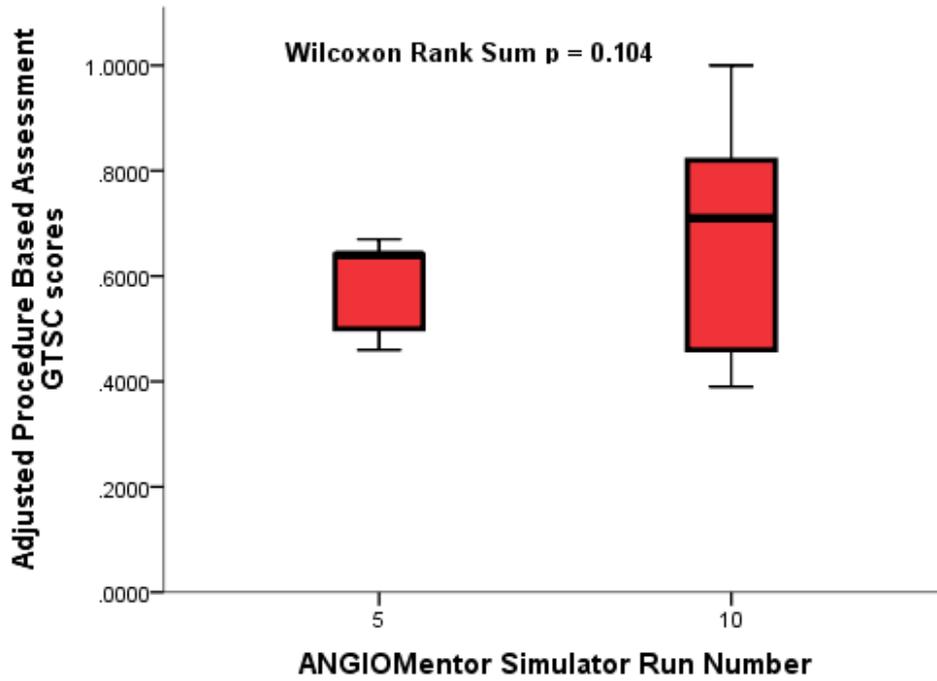
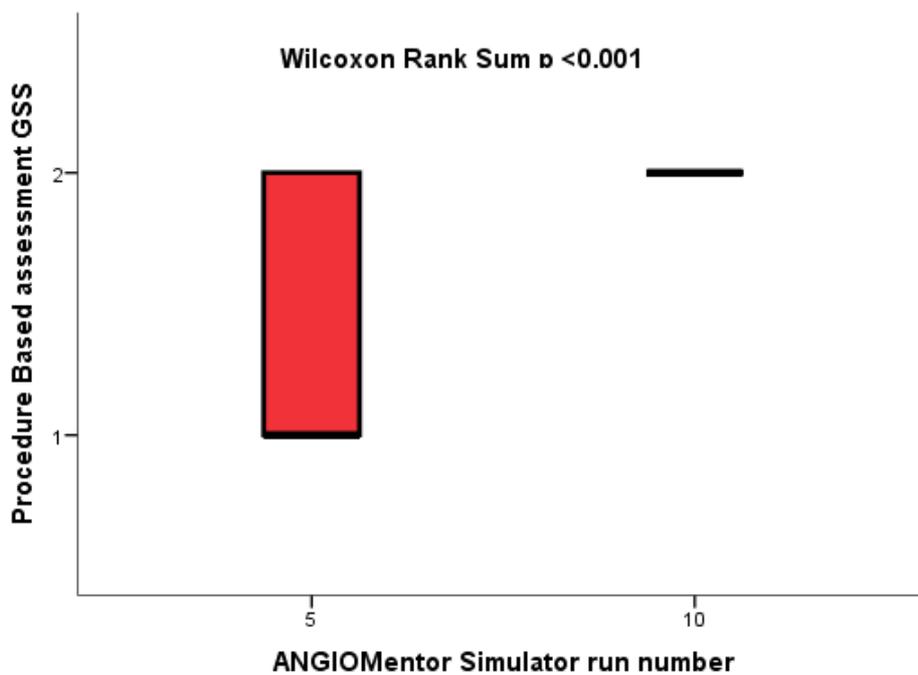


Figure 32 Box and whisker graph illustrating the cohort score for the PBA GSS between the 5<sup>th</sup> and 10<sup>th</sup> run on the ANGIOMentor simulator



### **3.5 Experiment 3- Intra-rater reliability of PBA**

#### **3.5.1 Participant recruitment and base line data**

Eleven individual raters were recruited to perform 2 individual assessments 6 weeks apart on the video which demonstrated a consultant performing a simulated sapheno-femoral junction dissection and ligation. Five consultants, 2 CT2, 2 ST4, one ST6 and 1 theatre nurse completed the first assessment. Seven individuals completed the second assessment with the target range of 6 weeks. Three further individuals completed the second assessment with 8 weeks of the first assessment (2 consultants and 1 trainee) and 1 CT2 failed to complete the second assessed video analysis. The data for all the individual assessment is demonstrated in table 13 (p 143). Only the 10 individuals who completed both assessments are included in the intra-rater reliability analysis.

#### **3.5.2 Results for Intra Rater Reliability**

The correlations for all 10 paired assessments are also described in table x. They indicate that the correlation between the 1<sup>st</sup> and second assessment is moderate only for the GTSC and the GSS and has not reached statistical significance. This is also demonstrated by Cronbach's Alpha reliability coefficients which are moderate at best. The relationships for the PBA GTSC and PBA GSS between the 1<sup>st</sup> and 2<sup>nd</sup> assessments are graphically illustrated below (figures 33-34 page 144).

**Table 13 Primary data from experiment 3 illustrating the scores provided by individual raters between their primary and secondary assessment of the simulated sapheno-femoral junction ligation.**

<b>N=10</b>	<b>Primary Assessment</b>		<b>Secondary Assessment</b>		<b>Spearman's rank correlation coefficient &amp; p values</b>
<b>Assessor designation</b>	Adjusted PBA GTSC score	PBA GSS	Adjusted PBA GTSC score	PBA GSS	PBA GTSC score r= 0.606 p= 0.064  Cronbach's Alpha=0.752
<b>Consultant</b>	0.53	2	0.50	3	PBA GSS r= 0.504 and p=0.137  Cronbach's Alpha =0.690
<b>Consultant</b>	0.58	2	0.33	2	
<b>Consultant</b>	0.52	3	0.75	3	
<b>Consultant</b>	0.69	3	0.55	2	
<b>Consultant</b>	0.50	2	0.44	3	
<b>ST4</b>	0.83	3	0.80	3	
<b>ST6</b>	0.69	3	0.67	3	
<b>CT2</b>	0.72	3	0.67	3	
<b>ST4</b>	0.72	4	0.78	4	
<b>Theatre nurse</b>	0.86	3	0.72	4	
<b>Medians (IQR)</b>	0.69 (0.53-0.78)	3 (2-3)	0.67 (0.485-0.758)	3 (2-3.25)	

Figure 33 Scatter graph illustrating Spearman's correlation between the PBA GTSC scores between the primary and secondary assessment of the simulated sapheno-femoral junction ligation

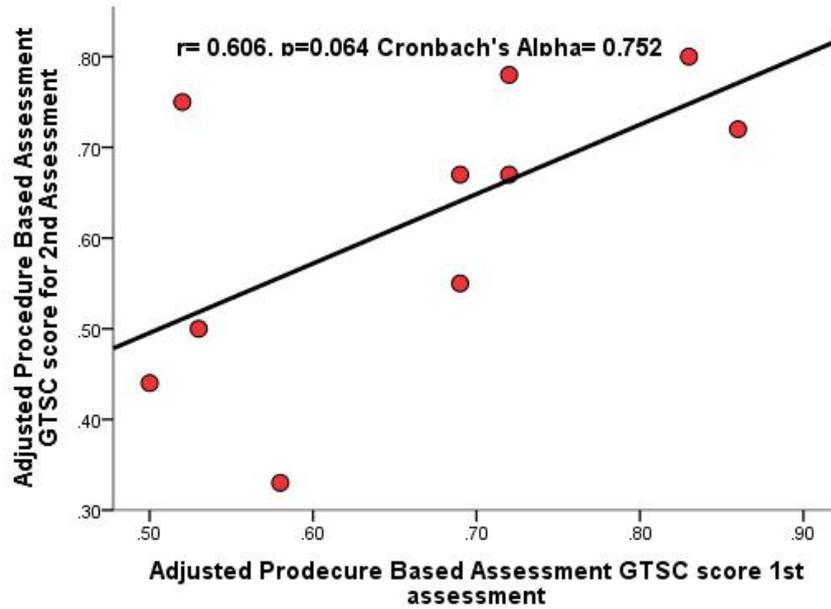
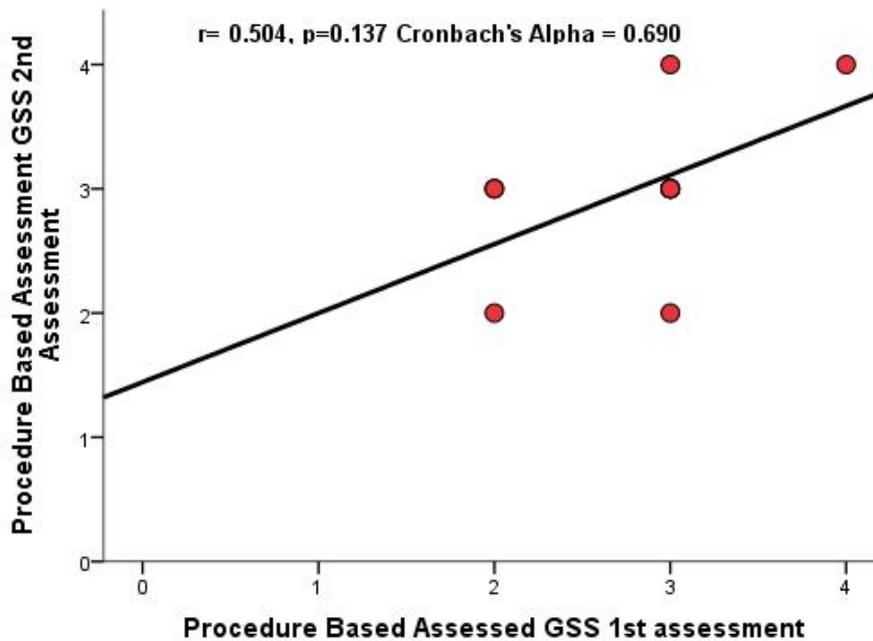


Figure 34 Scatter graph illustration the Spearman's correlation for the PBA GSS between the primary and secondary assessment of the simulated sapheno-femoral junction ligation



### 3.6 Results Summary

A modified PBA appears to be valid for aortic and sapheno-femoral junction ligation when compared to a trainees' actual operative experience in performing these procedures previously in the workplace. The global and task specific check list (GTSC) ( $r= 0.446$   $p= 0.029$ ) and global summary score (GSS) ( $r= 0.553$   $p= 0.005$ ) both demonstrated positive and statistically significant correlations with previous operative experience. However when compared to either just assisting or observing either sapheno-femoral junction or proximal aortic anastomosis PBA did not appear to be a valid measure of previous experience (GTSC  $r=0.294$   $p= 0.163$  and GSS  $r= 0.335$  and  $p=0.110$ ).

In contrast, the PBA tool was able to distinguish in part a moderate and statistically significant correlation with training grade. This was demonstrated by the GTSC ( $r= 0.586$   $p= 0.002$ ) but less convincingly by the global summary score, whose correlation values with training grade just missed statistical significance ( $r= 0.352$   $p= 0.080$ ).

The generic element of the objective structured assessment of technical skill (OSATS) demonstrated some interesting findings when compared to a modified PBA. It did not appear to demonstrate a statistically significant relationship between previous operative experience, either observation ( $r= 0.231$   $p= 0.302$ ) or performance ( $r= 0.347$   $p= 0.114$ ), but correlated highly with training grade ( $r= 0.634$   $p= 0.001$ ), outperforming the PBA in distinguishing between differing levels of trainee.

The checklist and generic element of PBA did not prove to be concurrently valid with the generic element of the OSATS tool. The Bland-Altman plot of methods of agreement between the two assessment methods

demonstrated that the results were statistically different from each other ( $p= 0.000$ ).

In evaluating inter-rater reliability; PBA was found to possess good to high statistical correlation between simultaneous raters during simulated vascular procedures for each distinct method. This was also re-enforced by a high Cronbach's reliability co-efficient (GTSC  $r= 0.655$   $p= 0.005$  Cronbach's Alpha 0.818 and GSS  $r= 0.843$   $p< 0.001$  Cronbach's Alpha 0.889) suggesting minimal elements of variance between raters in this particular arm of the study; however, estimation of intra-rater reliability in the final video study did not reach statistical significance. This was despite a moderate to good correlation co-efficient and a moderate to good Cronbach's reliability co-efficient. (GTSC  $r = 0.606$   $p= 0.064$  Cronbach's Alpha 0.752 and GSS  $r= 0.504$ ,  $p= 0.137$  Cronbach's Alpha 0.690)

Finally the PBA tool was demonstrated to be a responsive tool to repeated exposure and practice in performing SFA angioplasty, but this was only demonstrated in the global summary element (Wilcoxon- Rank Sum  $p= 0.001$ ) as opposed to the global and task specific checklist (Wilcoxon-Rank Sum  $p =0.104$ ). There was significant variation in individual performance, as highlighted in table 13 (page 143) and this result and others will be explored further in the following chapter.

## Chapter 4- Discussion

### 4.1 Chapter Overview

In this thesis the main aims were to determine if PBA possessed the psychometric properties of validity, reliability and responsiveness. In chapter 2 these properties were explored through a sequence of different experiments utilising vascular simulation. Specifically the main hypotheses were;

- Is PBA a valid measure of surgical performance as measured by previous surgical experience? Where previous surgical experience was determined by number of operations previously performed and observed and training level in vascular simulated procedures.
- Is PBA a reliable measure of surgical performance as determined by inter-rater and intra-rater reliability observing simulated vascular surgical procedures?
- Is PBA a responsive tool in measuring surgical performance as determined by observation of simulated surgical procedures?

In addition, data from experiments contained in this study allowed comparison of PBA with the well validated Objective Structured Assessment of Technical Skills (OSATS); in particular how it compared in terms of validity in relationship to previous operative experience, training level and whether PBA possessed concurrent validity with the OSATS tool.

The data from this study also provided evidence that the ANGIOMentor simulator was construct valid for differing levels of endovascular experience in novice, intermediate and expert participants performing a superficial femoral artery angioplasty. As there is little data in literature;

this study has provided evidence toward the continuing use of endovascular simulation in vascular training.

In chapter 3 the main findings of these studies were presented which illustrated that PBA appeared to possess some of these psychometric properties to a greater and lesser extent in simulated procedures. The following chapter will now explore each of those psychometric properties in detail, how these results fit in the context of the present literature in this area and what the implications of this research may be. Lastly this chapter will explore potential future directions of research in to PBA and simulation and whether this would prove to be a potentially useful area of research.

#### **4.2.1 Validity**

The definition of validity as discussed in chapter 1 is a broad church but ultimately can be defined by the statement. "does the test or evaluation measure what it purports to measure" Cronbach and Meehl (cited in Moss 1992)(53). The essence of construct validity also asked us to consider the directionality of the hypothesis' and whether this can be defined statistically.

In considering PBA as a tool to evaluate and measure surgical performance we sought to prove whether PBA was construct valid for two obvious variables which reflect individual surgical performance. These included the amount of previous exposure to a certain index procedures and an individual's total life time surgical exposure as measured by training grade/ year in training.

The main findings from experiment 1 produced very interesting results which may have implications for the assessment tool and how we train future surgeons. Firstly the finding that the index specific PBA tool was sensitive to the number of procedures previously performed, as opposed

previously observed suggests that the PBA is sensitive to increasing skill levels and repeated exposure and performance of index procedures. This was particularly reflected in the global summary score element of the PBA tool which demonstrates moderate correlation coefficients and a highly significant P values for numbers of procedures previously performed suggesting construct validity for the PBA tool. Contrasted to observation alone which was found to have low correlations and non-significant p-values in both elements of the PBA tool. This finding strengthens the evidence for PBA as a tool that can measure surgical performance (the essence of construct validity) and reflects the concept of deliberate practice (Ericsson(32, 184)) being the greatest contributor to developing expertise in a psychomotor skill domain. While there is little significant literature on PBA at present, the findings within this study also reflected those in Beard et al(58) where previous operative exposure was reflected in the assessment scores in the live operative environment. This result was also re-enforced in Sarker et al(96) who also demonstrated construct validity for the checklist element of the PBA form performing simulated laparoscopic cholecystectomy between two groups of experienced versus inexperienced operators. Howels et al(97) also indirectly demonstrated that PBA demonstrated construct validity for PBA during knee arthroscopy as those trainees who had been allowed simulated practice prior to performing a knee arthroscopy out performed those that no prior experience. Thereby demonstrating PBA possessed construct validity for previous experience.

The study finding also poses the question of how useful is observation only is in a training context, as this study demonstrated that the total number of index specific operations observed did not appear to correlate with performance? Clearly a certain degree of prior observation is required to performing a procedure, but after that it may only inform a trainee of those steps required to complete a procedure but does not contribute,

past a certain point, to the technical skill of that performance. After this repetition and muscle memory in conjunction with operative judgement must be built in order to improve technical performance. Where does the value of observation lie in a training context? Seeing others make mistakes or observation of different techniques to perform the same procedure would seem to have inherent educational benefit, but such non-tangible skills are not always directly measured by the didactic PBA form and fall into the non-technical skills such as decision making, and perhaps this also reflects that only the operative domains of the PBA form were used in this experiment. This strengthens the PBA tool as technical skill assessment, but perhaps not an assessment that can accurately assess those skills acquired through observation, such as decision making and situational awareness, which are often where surgical mistakes occur(185, 186).

The PBA scores were also correlated to training grade and a moderate and statistically significant relationship was demonstrated for the GTSC of the PBA. The GSS demonstrated a low correlation value and near statistically significant relationship for training grade. These results suggest partial construct validity for the PBA for training grade but are clearly not as robust as those for previous index specific performance. This may be because training grade represents a surrogate marker of technical ability for any given index procedures, i.e. trainees performing better due to their greater overall experience, although perhaps one might expect in that instance for the generic GSS to be more sensitive to this premise than the GTSC. However analysis of the OSATS result, where a purely generic assessment form was used identifies correlations that were very high and statistically significant for training grade but not statistically significant for number of procedures observed or performed, and may reflect the nature of the differences of the two forms used, one is very index specific and reflects actual technical skill in a particular procedure, whereas the OSATS tool used in this experiment is more generic and discriminates better

between overall generic skills which are accumulated with generic experience and increasing training grade.

Finally in attempting to describe a thorough assessment validity of the PBA tool we correlated the PBA assessment scores from experiment 1 with scores produced by the generic element of the OSATS assessment tool in the same experiment. This allowed determination of concurrent validity, another component of construct validity according to DeVon(48). There are no previous data which directly compares the PBA tool to the OSATS, while the two tools are marginally different one very index specific, the other very generic, they are attempting to measure the same construct; surgical technical ability. The OSATS tool has been well validated in numerous simulation and bench top studies(101, 102, 107, 113, 115) , However the Bland-Altman analysis which attempted to measure agreement between the two methods suggests poor concurrent validity in assessment of simulated vascular procedures. This finding may reflect the original design purpose of these two assessment tools. OSATS was originally designed for bench top procedures, whereas PBA was always designed to be used in the live workplace environment.

#### **4.2.2 Reliability**

##### **Inter-rater reliability**

It proved difficult to recruit sufficient individual raters to give inter-rater reliability data for the entirety of the bench top procedures from experiment 1 and only 3 of the bench top procedures were dually assessed, 2 sapheno-femoral junction ligations and 1 proximal aortic aneurysm anastomosis. However experiment 2 provided sufficient assessments by two or more raters for each assessed procedure to provide meaningful data which combined with the reliability data from experiment 1 allowed statistical analysis. Inter-rater reliability of both elements of the PBA tool proved to be moderate to high by calculation of correlation co-efficients

and reached statistical significance. The hypothesis free estimation of inter-rater reliability by Cronbachs alpha was extremely high for both elements of the PBA tool the GTSC and GSS reaching values of greater than 0.8 suggesting excellent reliability. This was especially so for the GSS element of the PBA tool. Beard et al(58) give a very thorough explanation for this effect. Their d-study of reliability of PBA in the workplace allowed estimation of the causes of variance amongst each element of the PBA tool and showed that such global scores anchored by explicated descriptors as found in the PBA GSS reduce variance amongst assessors and increase sensitivity to trainee ability; (the ideal characteristics of an assessment tool) a lesser effect is seen in the GTSC. Such global scores have been proved to be more valid and reliable than checklists in the literature (58, 102, 107).

Sarker et al(96) also assessed inter-rater reliability during live and simulated laparoscopic cholecystectomy performed by 28 trainees and found high reliability using the generic skill elements of the checklist part of the PBA form; with Kappa values of 0.86 for live and 0.84 for simulated laparoscopic cholecystectomy. However they did not assess the intra-rater reliability; the property of a tool to be consistent over time, this was also deliberately excluded from the methodology by beard et al(58) as they sought to assess trainees on the same index case as close together as possible to reduce a practice effect. In addition where video analysis was performed in order to assess trainees in a totally blinded manner, difficulties were found in editing and determining how much each trainee contributed to the case, and needed assistance, a concept that will be revisited again during this discussion.

### **Intra- rater reliability**

In contrast to the results for inter-rater reliability, intra-rater reliability, the estimate of reproducibility of the scores from the PBA at two different time points proved to be less robust. The correlation co-efficients for both elements were moderate at best and did not approach statistical significance; hypothesis free estimates of reliability of Cronbachs were moderate to good, between 0.69 for the GSS and 0.75 for the GTSC. Interestingly here the GTSC scores were more reliable than the generic global summary score as seen in the results from the inter-rater experiment. Why this should have occurred could be due to a number of reasons; the median and IQR for both first and second assessments were very similar and it may be that the small numbers of assessments resulted in the small differences in the first and second scores from each rater to be given too much weight within the reliability calculations (in effect a type 2 error, the result could be a false negative).

Why the GTSC should be demonstrated to be more reliable than the PBA GSS in contrast to the literature supporting higher reliability of generic assessment method may reflect the nature of the procedure being assessed. The GTSC approached a statistically significant result  $p= 0.064$  and may reflect that the videotaped procedure was a didactic step wise procedure, more easily reflected by a checklist type of assessment method, where the rater has minimal data on how much the assistant in helping or global skills being demonstrated. The video focuses only on the procedure being performed, there is little shown on instrument selection. There is no sound available to allow the assessor to determine the trainee's direction of the assistant which may make it more difficult to provide a consistent summary judgement. Interestingly the consultant scores on initial evaluation appear to show less consistency between the 1<sup>st</sup> and 2<sup>nd</sup> assessment. This may reflect the difficulty they found in determining strictly between the two definitions of "can do with minimal assistance"

and “can perform the procedure or part under supervision” for the global summary score reflecting the lower reliability of this element of the PBA tool in this case. The consultant assessors also show a general trend for lower marks than the trainees and other raters. Outside those studies specifically on PBA, some of the literature on surgical assessment that suggests that trainees are more likely to under or overestimate colleague performance(36, 187-189) there still appears to be less consistency between the first and second consultant assessment which in all respects should be similar as the procedure observed was exactly the same. We still have little data of the psychometric properties of PBA, especially in simulated procedures and many small studies which assess test- retest reliability (or intra-rater reliability) do not describe using use the same procedure with the same assessors to evaluate the assessment in question but rather use sequential videotaped performance (190, 191) or a combination of simulation and live procedures which of course includes an element of practice effect which would alter the results(191, 192). Such studies that do exist limit themselves to either construct validity and/ or inter-rater reliability of an instrument, rather than discovering if the assessment tool is reproducible over time(68).

Clearly there is considerable variation in the consultant scoring within this study which suggest in light of this example there is considerable day to-day variation in how consultant assessors perceive the simulated performance. While it could be easy to accept small differences between the GTSC hence the higher reliability co-efficients for these score between the first and second assessments, the difference between at GSS of 2 and 3 is fairly substantial in translation to whether an individual maybe competent or not and perhaps that is the limitation of the video and the nature of simulated assessments. The assessor is perhaps more likely to treat such global summary scores as a linear measurement rather than the anchored and explicit statement that they are meant to represent when

called to make a summary judgement on a real life theatre performance as they have very few external clues to the operator's global performance, without multiple camera angles and audio description(22, 58).

### **4.2.3 Responsiveness**

Experiment 2 was designed to highlight a crucial psychometric property of any assessment tool used to monitor performance; Responsiveness. Is the tool sensitive to changing levels of individual performance and practice over time? The main finding in this study shows that the PBA tool was sensitive to performance and this is highlighted in the GSS which showed a significant difference ( $p = <.001$ ) in the performance on the 5<sup>th</sup> and 10<sup>th</sup> runs on the ANGIOMentor simulator. This was not reflected in the GTSC check list scores which did not reach statistical significance.

Clearly each of the raters felt the majority of novice trainees had improved considerably or had retained a stable performance over time as all the trainees received a GSS level 2 on their second assessed run. Novice 1 and Novice 4 appeared to have stable performance, while novice 5 appeared to perform less well on the second run as demonstrated by low GTSC score, although had been given a better GSS than the first run.

Those novices who showed the greatest improvement between the 5<sup>th</sup> and 10<sup>th</sup> run on the simulator (novices 3, 4 and 7) were performed between 84 and 114 days apart. This included some of the longest intervals between assessments and suggests that either that skills performance can be preserved within this length of time, or that this may be a way of identifying individuals who naturally retain psychometric performance better despite the longer reassessment window thus may be a useful method of selecting trainees who are more suited to endovascular work.

However there appears to no preset defined interval for skills retention, Ericsson suggests that 6 months is the maximum amount of time before

skill degrades appreciably(32), however intervals from 4 months to one year have been studied for both laparoscopic(193) and basic surgical skill(194) and have demonstrated either no or very little degradation of skill via objective assessment. Indeed, when this is also been objectively assessed on intensive simulated skill courses, such as those now being integrated into Canadian training; after six months, high retention rates were demonstrated(137). All our novices completed the final assessments well within this time period.

There have been no studies on the ANGIOMentor or PBA that have looked at whether it is a responsive tool for individual trainees. The ANGIOMentor has been proved to be construct valid in a some studies where increasing experience has be associated with increased scores(195, 196), but these studies have used single assessments of trainees or students at different stages of training and while encouraging is not strictly a surrogate assessment of responsiveness. Certainly this study shows that there are elements of the PBA form that are responsive to a practice effect, a desirable characteristic in an assessment tool. Once again, like many educational and indeed endovascular studies the small numbers of participants can limit the generalizability of the results and there have been at least 2 novices been assessed by only one rater, which given the overall small numbers of assessments could of skewed the data. It proved extremely difficult to get sufficient raters for each of the assessments and the difficulty in organising this is reflected in the absence of a second rater from some of the initial assessments and the length of time taken for some of the novices to undergo their second assessment.

Studies that evaluate skills retention suggest that distributed practice, rather than massed practice are more likely to result in greater skills retention(32, 197), contrary to this particular study design. But the lack of sufficiently qualified raters did limit the timing of the practice sessions.

Factors that can also effect skill retention during simulated procedures can include; trainee motivation, difficulty of the task and complexity of the procedure. Level of integration within the simulation such as whether it is physical verses cognitive in nature and need for speed and accuracy can also affect retention. Skill retention is also affected by whether the task is natural as opposed to artificial and if the task is well defined as opposed to one which requires a continuous response (Arthurs et al in Lammers et al) (198). As some of our novices were research nurses as opposed to surgical trainees, with a vested interest in learning on the endovascular simulator, there may have been differing levels of motivation to perform on the simulator. The step wise protocol created for the novices to use as a guide was complicated, and some novices may have been able to integrate the cognitive and physical tasks of the procedure better than others, which improved their GTSC performance. Whereas others found the individual steps more difficult to retain, but the familiarity with the simulator and handling the wires improved their overall performance from baseline as demonstrated by the improvements in the GSS from the first base line measurement. The task was well defined with clear objectives but there were many steps and the complexity to the task did mean that there many opportunities for trainees to go astray.

### **4.3 Limitations of this research**

With the exception of experiment 3 where the performance on the video assessment was anonymised all the other experiments were conducted in an un-blinded manner. This is obviously is a less robust study design but reflects how the PBA form is presently used in the work place, by trainers who know their trainee. However as experiment 1 was conducted at the royal college of surgeons the majority of trainees were unknown to each other and the designated assessors and met for the first time on the course. In experiment 2 while some of the trainees were known to the assessors, more than 50% of the assessments were performed by

interventional radiologists who had never met the novices before in the clinical environment so providing a more objective assessment of performance.

The small number of participants within these studies reflects many interventional studies and trainee studies where live raters are required to assess trainees. This reflects the difficulty that many educational researchers encounter in gathering both raters and trainees who are in busy clinical practice. However it was important to demonstrate that the PBA tool could be used in the very population who would normally be using it in clinical practice, rather than students, who may have been more readily available. It was also very important to use live raters who can more readily assess global performance and how trainees actually respond when being assessed. As Beard et al(58) demonstrated, each element of the assessment including the rater assessment of the case difficulty, assessor stringency and assessor designation stringency can contribute to variance. That any assessment tool possesses these variances is likely inherent and difficult to remove entirely, but there are few assessment methods than can pragmatically be used in clinical practice, outside the research environments and the majority should be used with a rater because trainee performance is improved considerably with specific relevant expert feedback(34-36, 184, 199).

Of particular note and highlighted in the graphical illustration of the validity results in experiment 1 (p119-126) is clear evidence of data clustering in terms of operative experience which in conjunction with the small numbers of trainees and could of skewed the statistical analysis. This is in contrast to training grades as illustrated in the figures 16,17 and 20, which show an more even distribution of data points. Unfortunately this may be a reflection of the study design where recruitment was governed by the attendance at the training course. Inherent in this is the fact that the

majority of attendees had little or moderate experience in the procedures analysed here which was why they were attending the course and were more likely to be junior trainees.

These studies do not add to the evaluation of the PBA as a tool to provide formative or summative feedback, they have only looked at the psychometric properties of PBA, which suggest they would be unsuitable for summative evaluation as only partial reliability has been demonstrated. They have been confined to a snap shot assessment. Within the utility formula outlined in the chapter one only validity and reliability have been explored, acceptability, cost effectiveness and educational impact have not been included in this study. However that PBA possesses validity in vascular simulation, particularly for previous operative experience and has and has been shown to be responsive to practice effect; a surrogate for educational impact, is encouraging data for its use in evaluating trainee performance in simulated practice.

#### **4.4 Implications of this research**

That the PBA tool has to been found to be statistically valid, possesses inter-rater reliability and has been found to be in part responsive in assessing simulated vascular procedures is encouraging, but true validity should include within the definition the use to which the tool is put. If PBA can be used to assess simulated procedures there is no evidence as yet to suggest that this tool should be used in anything but a formative manner to assess trainees in simulated procedure and perhaps its use as a benchmarking tool in the ARCP process for workplace assessment is unjustified beyond that of formative assessment, certainly when it has been used for simulated assessments. There has been no evidence yet to suggest that it could be used to certify trainees as competent using simulation even in the small number of limited studies on PBA, where

reliability values are at or around 0.8 at best and reliability co-efficients of > 0.9 are preferred for such high stakes assessment(43).

What this study does highlight is that PBA is a feasible tool in assessing trainees in a curriculum which has a growing element of simulation. It seems inherently practical to continue to use a tool which is already in clinical practise and can potentially seamlessly continue simulation training into the real world environment. The majority of raters using the PBA form found it easy to use and trainees and trainers are now becoming more familiar with its use suggesting that acceptability is less of a problem than it has been in the past.

Highlighted in this study was also the appearing procedure specific nature of the PBA tool. It was clearly sensitive to a trainee's previous operative performance and exposure to index procedures, far more so than the generic OSATS tool in these vascular simulation studies. Worryingly over the past 2 years there have been changes on the ISCP website which have started comparing an individual's performance with cohort level performance in workplace based assessments. This position is not justified based on the evidence, especially given that PBA is a very index specific procedure and may not fully represent a trainee's generic skill level, and so may reflect unfavourably on a trainee, if they have limited exposure of a certain speciality or procedure.

#### **4.5 Future directions**

This study has been able to demonstrate that PBA is a valid measure of surgical skill as defined by previous operative experience in a limited number of simulated operations. To be a truly valid assessment, certainly in vascular simulation, it should be validated with the entire range of operations available on the ISCP website, where simulation opportunities for training are available. This could be achieved in part by simulated cadaveric operations with a range of trainees of differing abilities, or high

fidelity mock theatre environments to replicate entire operations such as large open operations where only low fidelity jigs and resources are otherwise available. This is potentially an important next step in vascular curriculum development as integrating simulation at all levels of vascular training will become more commonplace and a valid and reliable method of national assessment needs to be agreed upon if simulation is going to be an accredited part of the new vascular training curriculum.

That intra-rater reliability was not fully demonstrated in this study was disappointing, although correlation co-efficients and hypothesis free estimates of reliability were moderate they did not reach statistical significance. I believe it would be useful to repeat this experiment with a more sophisticated video analysis of performance, with additional camera angles and sound to allow the raters to gain a better overall impression of the individual performance. Alternatively if these resources were not available the PBA form could be edited to only allow demonstration of those competencies that could be demonstrated by a single camera angle. A greater number of raters would also potentially reduce any type 2 errors and increase the robustness of the data; potentially sending the video to other experienced raters in additional hospitals would reduce any potential bias and may result in a more convincingly positive result.

In addition to exploring validity and reliability it would complete the global assessment of PBA by exploring the additional concepts defined in the Utility Formula, acceptability, cost effectiveness and educational benefit. Utilising a randomised study design trainees could be assigned to received feedback, either using the structured PBA tool and verbal intra-operative feedback or via verbal feedback alone, then assessed again performing the same procedure by a rater blinded to the type of feedback. An endovascular simulator procedure might lend its self well to such a study design and demonstrate well if PBA adds to the educational benefit of

expert feedback. Acceptability could also be evaluated at the same time. This concept is particularly important as many trainers and trainees still report finding that completing PBA forms is tedious and add little to educational encounters. The PBA form received a lot of negative literature when it initially was introduced in to clinical practice(118, 200) and in order to complete the entirety of the PBA was time consuming from the consultant assessor point of view, particularly if they assessed each checklist competency by direct observation of the trainee(119). In complex operations this can be in excess of 45 individual competencies, including consent, pre-operative planning and preparation, exposure and closure, intra-operative domains, and post operative management. Additionally there are still many lost opportunities for PBA to be performed in the workplace(65). It would be useful to assess whether the initial views of many had changed since the early years of its introduction. Indeed Beard et al attempted to assess the acceptability of PBA during a recent workplace based study(58) and found that the both assessors and trainees reported subjectively moderate to good reports on ability to improve feedback and improve performance. This unfortunately was not a comparative or outcome based element of the study which compared PBA to verbal feedback alone and relied on a structured questionnaire data that did not include negative statements or free text responses i.e. qualitative data which may have added a useful component to the acceptability data.

One area in which the acceptability and potential educational benefit of PBA could be improved is in the re-design and expansion of the global summary score. At present while it provides a good reliability component to the PBA, it may not track a trainee's progress well through a surgical placement i.e. be responsive, and a proposed change to the explicit descriptors in the global summary score may improve this(201). (Appendix 7)

Cost effectiveness of PBA may be impossible to prove since improved educational outcomes may result in many tangible economic benefits, such as reduced number of complications, decreased hospital stay and reduced litigation. However there is an appreciable cost associated with maintaining this assessment system partially funded by the trainees that use it. Therefore it could equally be argued that the cost of running and maintaining this method of curricular assessment may not justify the results, unless these tangible benefits can be demonstrated; an immensely difficult task.

Where there is high internal consistency between two such components as a checklist and generic summary score or generic scoring system such is found in PBA or the OSATS tool, and where such checklist are found to be less reliable than the generic scoring systems, it may be possible to reduce itemised checklists to their least parts using a statistical method known as rotational factor analysis. Where related variables in a such a checklist can be analysed to examine where true variance in the related variables lie, often reducing the number of variables to only those that provide significant variance in the result can result in a less complex, unwieldy and more easily filled in form or questionnaire, which may be a desirable characteristic, when ease of use and acceptability are being analysed. This potentially may make it more suitable for simulation assessment, but could reduce the educational impact by reducing the potential opportunities for feedback.

Ultimately if such radical changes to the PBA were made, then the purpose of PBA must be made explicit. It appears now that specific numbers are a requirement for progression through a satisfactory Annual Review of Competency Progression each year for all surgical trainees (a summative approach), and yet its original purpose was to provide trainees with feedback and assessment of learning (a formative approach). While factor

analysis can strengthen the validity and reliability and acceptability of PBA, it could be at a loss of the educational benefit? Therefore perhaps one of the first steps in further rationalising and streamlining any assessment system would be to first establish this fact; otherwise the whole ethos behind the workplace based assessment system and individualised competency based progression is flawed.

As touched upon earlier, the ISCP website has now started comparing trainee performance at each training level to the national averages, that it has contained within its database, a rather worrying trend for a curriculum that advocates a competency based approach centred around the individual trainee, and is moving the PBA from away from its formative beginnings to a summative assessment or credentialing assessment.

If this is to be the case then the data that already exists on PBA should be used in rationalised manner, rather than a scatter gun approach. Consider then, that the evidence from studies such as this one, provide evidence that PBA can be a valid and potentially reliable measure of performance in simulation training. This can provide feedback in a formative manner to aid trainee progression and learning. Such studies as Beard et al(58) can provide data to support a rationalised number of assessments per trainee, per index procedure in the workplace, in essence providing a robust, valid, reliable assessment of performance which is evidence based.

Indeed this difficulty in defining the purpose of WBA's as tools to enable feedback verses an assessment tool which provides evidence of progress has been recognised by many groups and the GMC. Discussion has taken place regarding the introduction of an assessment process known as a Supervised Learning Events (SLE's), designed to give formative feedback only. This would be contrasted with an Assessment of Performance (AoPs) which would provide summative evidence of progress. Quite rightly this proposal has highlighted that once a trainee has achieved a certain

standard, as bench marked against the curriculum, only a minimal number of AoP should need to be performed to inform progression. This recognises that good evidence now exists for rationalising the numbers of WBA especially PBA in the work place(202).

Informed by ideas from this work and others I would propose that a DOPS as they presently exist be discarded as there is not useful evidence for their continued use and introduce a generic formative assessment tool which could be used throughout the continuum of surgical training, simulation and theatre based, from CT1 to CT8 (appendix 7). This is an amalgamation from the generic elements of an OSATS tool and the GSS from PBA as informed by the proposed changes to the PBA GSS to make it more representative of real world practice.

The generic section of the OSATS tool in this study was the most valid measure of training grade and marker of generic skill, whereas the modified PBA proved to be more valid for index specific performance especially the checklist within the context of this study. The PBA GSS was found to possess high inter-rater reliability within the context of this study and also was the most reliable component of the PBA tool in the study by Beard et al(58) therefore the obvious implication of this is to use the OSATS to formatively assess generic skill, a SLE (Supervised Learning Event), with the PBA GSS to provide a reliability element. Whereas the checklist elements combined with the proposed changes to the PBA GSS could then be used to provide a summative assessment or AoP (Assessment of Performance) for skill for all index procedures from ST5 onwards when trainees enter surgical sub- speciality training. Each index procedure, for each speciality, for each training level is already defined on the ISCP website and so trainees are then informed what they need to achieve for each level of training, with the minimum amount of paperwork.

The benefits of using the generic OSTATS tool for formative assessment allow both the trainee and trainer to demonstrate tangible improvements in the trainee's performance which is translated to a linear numerical measurement which can plot their progress during placements and on a yearly basis while still being able to provide specific feedback for generic skills improvement.

#### **4.6 Dissemination of work**

The work undertaken as been presented in part in poster format at the Vascular Society of Great Britain and Northern Ireland meeting in Edinburgh and the Association of surgeons in Training meeting in Cardiff. Elements have also been presented orally at the Annual meeting of Faculty of Surgical Trainers, Birmingham 2012. There are presently two papers based on this work awaiting approval for publication and it is hoped that that Joint Committee on Surgical Training (JCST) Assessment Group who are presently considering how simulation will fit in to the surgical curriculum will find the evidence in this thesis a useful contribution to their work. The JCST strategic aims 2013-2018 include a commitment to a "curriculum which moves with new developments and has an active role in educational research and developing an effective assessment system"(203).

#### **4.7 Conclusion**

The experiments performed in this study have demonstrated that modified PBA is a valid instrument with which to measure surgical performance in vascular simulated procedures as it appears to be sensitive to previous primary operative experience. It has not been demonstrated to have concurrent validity with the "gold standard" measure of bench top and

simulated surgical performance, the Objective Structured Assessment of Surgical skill; reflecting perhaps, the differing origin of purpose of PBA and OSATS. However it has been shown to be a responsive instrument in endovascular simulation assessment. This suggests its use in simulated vascular assessment can be justified.

However, only partial reliability has been demonstrated in these studies. While possessing high inter-rater reliability, intra-rater reliability was moderate only and did not reach statistical significance. This implies that PBA should only be used for formative skills assessment in simulated vascular procedures and further work on reliability should be undertaken if PBA was to be considered for summative simulation assessment in the future.

While PBA has both its advocates and detractors it clearly has a role in formative assessment in simulation and in the workplace and this study has added to the presently small volume of literature on PBA. As the evidence base for the PBA tool increases in a wider number of procedures hopefully it can be utilised to rationalise, simplify and streamline the assessment tool and process for trainers and trainees alike.

## References

1. Carraccio C, Wolfsthal SD, Englander R, Ferentz K, Martin C. Shifting Paradigms: From Flexner to Competencies. *Academic Medicine*. 2002;77(5):361-7.
2. Wilbrink B. Assessment in historical perspective. *Studies in Educational Evaluation*. 1997;23(1):31-48.
3. Bowker L, Hand C, Young R. Clinical teaching: past, present and future. In: Penelope Cavenagh, Leinster SJ, Miles S, editors. *The Changing Face of Medical Education*: Radcliff Publishing; 2011. p. 79-90.
4. ISCP. [www.iscp.ac.uk/home/assessment\\_overview.aspx](http://www.iscp.ac.uk/home/assessment_overview.aspx). Intercollegiate Surgical Curriculum Project; 2012 [cited 2012 14th Feb].
5. Memon MA, Brigden D, Subramanya MS, Memon B. Assessing the surgeon's technical skills: analysis of the available tools. *Acad Med*. 2010 May;85(5):869-80.
6. Moorthy K, Munz Y, Sarker SK, Darzi A. objective assessment of technical skills in surgery. *bmj*. [review]. 2003 1/11/2003;327:1032-7.
7. Darzi A, Mackay S. Assessment of surgical competence. *Qual Health Care*. 2001 Dec;10 Suppl 2:ii64-9.
8. Prideaux D. The emperor's new clothes: from objectives to outcomes. *Med Educ*. 2000 Mar;34(3):168-9.
9. Harden RM. Developments in outcome-based education. *Med Teach*. 2002 Mar;24(2):117-20.
10. Newble D, Stark P, Bax N, Lawson M. Developing an outcome-focused core curriculum. *Med Educ*. 2005 Jul;39(7):680-7.
11. Smith J. *The Shipman Inquiry*2003.
12. Kennedy I. *The Inquiry into the management of care of children receiving complex heart surgery at the Bristol Royal Infirmary*2001.
13. Greenhalgh T, Wong G. Revalidation: a critical perspective. *Br J Gen Pract*. 2011 Mar;61(584):166-8.
14. Isles M P. Surgical procedure-based assessment: How did we get here and where are we going? *Annals of the Royal College of Surgeons of England (Suppl)*. 2010;92(7):1-3.
15. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *BMJ*. 2010;341.
16. Temple J. *Time for Training: A review of the Impact of the European Working Time Directive on the Quality of Training: Medical Education England and Department of Health*2010.
17. Graham D, Thompson A, Gregory S. *Liberating Learning: COPMeD: Conference Of Postgraduate Medical Deans 2009 2002 (Revised 2009)*.
18. Kneebone RL, Scott W, Darzi A, Horrocks M. Simulation and clinical practice: strengthening the relationship. *Medical Education*. 2004;38(10):1095-102.
19. Bismuth J, Donovan MA, O'Malley MK, El Sayed HF, Naoum JJ, Peden EK, et al. Incorporating simulation in vascular surgery education. *J Vasc Surg*. 2010 Oct;52(4):1072-80.
20. Jakimowicz J, Fingerhut A. Simulation in surgery. *British Journal of Surgery*. 2009;96(6):563-4.

21. Gould D. Using simulation for interventional radiology training. *British Journal of Radiology*. 2010 July 1, 2010;83(991):546-53.
22. Duncan JR, Glaiberman CB. Analysis of Simulated Angiographic Procedures: Part 1--Capture and Presentation of Audio and Video Recordings. *Journal of Vascular and Interventional Radiology*. 2006;17(12):1979-89.
23. Lloyd GM, Maxwell-Armstrong C, Acheson AG. Fresh frozen cadavers: an under-utilized resource in laparoscopic colorectal training in the United Kingdom. *Colorectal Dis*. 2011 Sep;13(9):e303-4.
24. Nugent E, Hseino H, Boyle E, Mehigan B, Ryan K, Traynor O, et al. Assessment of the role of aptitude in the acquisition of advanced laparoscopic surgical skill sets : Results from a virtual reality-based laparoscopic colectomy training programme. *Int J Colorectal Dis*. 2012 Apr 17.
25. Aggarwal R, Grantcharov T, Moorthy K, Hance J, Darzi A. A competency-based virtual reality training curriculum for the acquisition of laparoscopic psychomotor skill. *Am J Surg*. 2006 Jan;191(1):128-33.
26. Gurusamy KS AR, Palanivelu L, Davidson BR. Virtual reality training for surgical trainees in laparoscopic surgery. *Cochrane Database of Systematic Reviews*. 2009(Issue 1):Art No: CD006575.
27. Torkington J, Smith SG, Rees BI, Darzi A. Skill transfer from virtual reality to a real laparoscopic task. *Surg Endosc*. 2001 Oct;15(10):1076-9.
28. Torkington J, Smith SG, Rees B, Darzi A. The role of the basic surgical skills course in the acquisition and retention of laparoscopic skill. *Surg Endosc*. 2001 Oct;15(10):1071-5.
29. Sturm LP, Windsor JA, Cosman PH, Cregan P, Hewett PJ, Maddern GJ. A systematic review of skills transfer after surgical simulation training. *Annals Of Surgery*. 2008;248(2):166-79.
30. Crochet P, Aggarwal R, Dubb SS, Ziprin P, Rajaretnam N, Grantcharov T, et al. Deliberate practice on a virtual reality laparoscopic simulator enhances the quality of surgical technical skills. *Ann Surg*. 2011 Jun;253(6):1216-22.
31. Barry Issenberg S, Mcgaghie WC, Petrusa ER, Lee Gordon D, Scalese RJ. Features and uses of high-fidelity medical simulations that lead to effective learning: a BEME systematic review\*. *Medical Teacher*. 2005;27(1):10-28.
32. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. *Acad Med*. 2004 Oct;79(10 Suppl):S70-81.
33. Reznick RK, MacRae H. Teaching surgical skills--changes in the wind. *N Engl J Med*. 2006 Dec 21;355(25):2664-9.
34. Grantcharov TP SS, Kristiansen V B. The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. *Surg Endosc*. 2007;21:2240-3.
35. Porte MC, Xeroulis George, Reznick R K, Dubrowski. Verbal feedback from an expert is more effective than self-accessed feedback about motion efficiency in learning new surgical skills. *The American Journal of Surgery*. 2007;193:105-10.
36. Pandey VA, Wolfe JHN, Black SA, Cairls M, Liapis C, Bergqvist D. Self-assessment of technical skill in surgery: The need for expert feedback. *Annals of the Royal College of Surgeons of England*. 2008;90(4):286-90.
37. McGaghie WC, Issenberg SB, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. *Med Educ*. 2010 Jan;44(1):50-63.

38. Boyle E, O'Keeffe DA, Naughton PA, Hill ADK, McDonnell CO, Moneley D. The importance of expert feedback during endovascular simulator training. *Journal of Vascular Surgery*. 2011;54(1):240-8.e1.
39. Michelson JD, Manning L. Competency assessment in simulation-based procedural education. *Am J Surg*. 2008 Oct;196(4):609-15.
40. Cook DA, Beckman TJ. Current concepts in validity and reliability for psychometric instruments: theory and application. *Am J Med*. 2006 Feb;119(2):166 e7-16.
41. Ludlow HH. GALTON: the first psychometrician. *Popular Measurement*. 1998;Spring:13-4.
42. Jones LV, Thissen D. A history and overview of psychometrics. In: Rao CR, Sinharay S, editors. *Handbook of statistics 26: Psychometrics*. Amsterdam: Elsevier; 2007. p. 1-22.
43. Chandratilake M, Davis M, G P. Evaluating and designing instruments for medical education: the utility formula. *The Internet Journal of Medical Education [serial on the Internet]*. 2010; 1(1).
44. Van Der Vleuten CPM. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education*. 1996 1996/01/01;1(1):41-67.
45. Roach KE. Measurement of Health Outcomes: Reliability, Validity and Responsiveness. *JPO: Journal of Prosthetics and Orthotics*. 2006;18(6):P8-P12.
46. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. *J Clin Epidemiol*. 2000 May;53(5):459-68.
47. S B Issenberg, McGaghie WC, Petrusa ER, Gordon DL, Scalese RJ. What are the features and uses of high-fidelity medical simulation that lead to most effective learning? BEME Guide No 4. *Medical Teacher*. 2005;27(1):10-28.
48. DeVon HA, Block ME, Moyle-Wright P, Ernst DM, Hayden SJ, Lazzara DJ, et al. A Psychometric Toolbox for Testing Validity and Reliability. *Journal of Nursing Scholarship*. 2007;39(2):155-64.
49. Kline TJB. *Classical test theory. Assumptions, Equations, Limitations and Item Analysis*. Psychological Testing: SAGE publications; 2005. p. 91-106.
50. *Developing and Maintaining an Assessment System- a PMETB guide to good practice* 2007 January 2007.
51. Van Nortwick SS, Lendvay TS, Jenson AR, Wright AS, K.D. H, S. K. Methodologies for establishing validity in surgical simulation studies. *Surgery*. [review]. 2010;147(5):622-30.
52. Colliver JA, Conlee MJ, Verhulst SJ. From test validity to construct validity ... and back? *Med Educ*. 2012 Apr;46(4):366-71.
53. Moss PA. Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment. *Review of Educational Research*. 1992 Fall 1992;62(3):229-58.
54. Borsboom D, Mellenbergh GJ, van Heerden J. The concept of validity. *Psychol Rev*. 2004 Oct;111(4):1061-71.
55. Messick S. Validity in Psychological Assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*. 1996;50(9):741-9.
56. Downing SM. Validity: on meaningful interpretation of assessment data. *Med Educ*. 2003 Sep;37(9):830-7.

57. Lawshe CH. A Quantitative approach to content validity. *Personnel Psychology*. 1975;28:563-75.
58. Beard JD, Marriott J, Purdie H, Crossley J. Assessing the surgical skills of trainees in the operating theatre: a prospective observational study of the methodology. *Health Technology Assessment*. [Research Support, Non-U.S. Gov't]. 2011 Jan;15(1):i-xxi, 1-162.
59. Waltz CF, Strickland OL, E.R.Lenz. *Measurement in nursing and Health Research*. 3rd ed. New York: Springer; 2005.
60. Gagnon MP, Duplantie J, Fortin JP, Jennett P, Scott R. A survey in Alberta and Quebec of the telehealth applications that physicians need. *J Telemed Telecare*. 2007;13(7):352-6.
61. NHS Management Executive. *Junior Doctors: The New Deal*. (Department of Health)1991.
62. Beard JD. Assessment of surgical skills of trainees in the UK. *Annals of the Royal College of Surgeons of England*. 2008;90(4):282-5.
63. Beard J, Rowley D, Bussey M, Pitts D. Workplace-based assessment: assessing technical skill throughout the continuum of surgical training. *ANZ Journal of Surgery*. 2009;79(3):148-53.
64. Donaldson SL. *Unfinished Business: Proposals for reform of the house officer grade*. In: Health Do, editor.2002.
65. Marriott JC PH, Millem A, Beard JD. The lost opportunities for surgical training in the NHS. *Ann R Col Surg Engl (Suppl)*. 2011;93:202-6.
66. Beard JD. Assessment of surgical competence. *British Journal of Surgery*. 2007;94(11):1315-6.
67. Okoro T, Sirianni C, Brigden D. The Concept of Surgical Assessment: Part 2 Available Tools. *Bulletin of the Royal College of Surgeons of England*. 2010;92(9):324-6.
68. van Hove PD, Tuijthof GJ, Verdaasdonk EG, Stassen LP, Dankelman J. Objective assessment of technical surgical skills. *British Journal of Surgery*. 2010;97(7):972-87.
69. Ahmed K, Miskovic D, Darzi A, Athanasiou T, Hanna GB. Observational tools for assessment of procedural skills: a systematic review. *American Journal of Surgery*. 2011;202(4):469-80.e6.
70. Stylopoulos N, Vosburgh KG. Assessing technical skill in surgery and endoscopy: a set of metrics and an algorithm (C-PASS) to assess skills in surgical and endoscopic procedures. *Surg Innov*. 2007 Jun;14(2):113-21.
71. Gallagher AG, Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. Learning curves and reliability measures. *Surg Endosc*. 2002 Dec;16(12):1746-52.
72. Datta V, Mackay S, Mandalia M, Darzi A. The use of electromagnetic motion tracking analysis to objectively measure open surgical skill in the laboratory-based model. *J Am Coll Surg*. 2001 Nov;193(5):479-85.
73. Smith SG, Torkington J, Brown TJ, Taffinder NJ, Darzi A. Motion analysis. *Surg Endosc*. 2002 Apr;16(4):640-5.
74. Taffinder N, Smith S, Mair J, Russell R, Darzi A. Can a computer measure surgical precision? Reliability, validity and feasibility of the ISCAD. *Surg Endosc* 1999;13 (Suppl):81.
75. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Pappasavas P, Dosis A, et al. An Evaluation of the Feasibility, Validity, and Reliability of Laparoscopic Skills

Assessment in the Operating Room. *Annals Of Surgery*. 2007;245(6):992-9  
10.1097/01.sla.0000262780.17950.e5.

76. Francis NK, Hanna GB, Cuschieri A. The performance of master surgeons on the Advanced Dundee Endoscopic Psychomotor Tester: contrast validity study. *Arch Surg*. 2002 Jul;137(7):841-4.

77. Francis NK, Hanna GB, Cresswell AB, Carter FJ, Cuschieri A. The performance of master surgeons on standard aptitude testing. *Am J Surg*. 2001 Jul;182(1):30-3.

78. Francis NK, Hanna GB, Cuschieri A. Reliability of the Advanced Dundee Endoscopic Psychomotor Tester for bimanual tasks. *Arch Surg*. 2001 Jan;136(1):40-3.

79. Broe D, Ridgway PF, Johnson S, Tierney S, Conlon KC. Construct validation of a novel hybrid surgical simulator. *Surg Endosc*. 2006 Jun;20(6):900-4.

80. Egi H, Okajima M, Yoshimitsu M, Ikeda S, Miyata Y, Masugami H, et al. Objective assessment of endoscopic surgical skills by analyzing direction-dependent dexterity using the Hiroshima University Endoscopic Surgical Assessment Device (HUESAD). *Surg Today*. 2008;38(8):705-10.

81. Egi H, Okajima M, Kawahara T, Yoshimitsu M, Sumitani D, Tokunaga M, et al. Scientific assessment of endoscopic surgical skills. *Minimally Invasive Therapy & Allied Technologies*. 2010;19(1):30-4.

82. Tokunaga M, Egi H, Hattori M, Yoshimitsu M, Sumitani D, Kawahara T, et al. Approaching time is important for assessment of endoscopic surgical skills. *Minimally Invasive Therapy & Allied Technologies*. 2012;21(3):142-9.

83. Egi H, Tokunaga M, Hattori M, Ohdan H. Evaluating the correlation between the HUESAD and OSATS scores: Concurrent validity study. *Minimally Invasive Therapy & Allied Technologies*. 2013;22(3):144-9.

84. Chmarra MK, Bakker NH, Grimbergen CA, Dankelman J. TrEndo, a device for tracking minimally invasive surgical

instruments in training setups. *Sensors and Actuators A*. 2006;126:328-34.

85. Van Empel PJ, Van Rijssen LB, Commandeur JP, Verdam MG, Huirne JA, F S, et al. Validation of a new box trainer-related tracking device: the TrEndo. *Surg Endosc*. 2012;26(8):2346-52.

86. Vassiliou MC, Feldman LS, Andrew CG, Bergman S, Leffondre K, Stanbridge D, et al. A global assessment tool for evaluation of intraoperative laparoscopic skills. *Am J Surg*. 2005 Jul;190(1):107-13.

87. Gumbs AA, Hogle NJ, Fowler DL. Evaluation of resident laparoscopic performance using global operative assessment of laparoscopic skills. *J Am Coll Surg*. 2007 Feb;204(2):308-13.

88. Sroka G, Feldman LS, Vassiliou MC, Kaneva PA, Fayed R, Fried GM. Fundamentals of Laparoscopic Surgery simulator training to proficiency improves laparoscopic performance in the operating room—a randomized controlled trial. *American Journal of Surgery*. 2010;199(1):115-20.

89. Tang B, Hanna GB, Carter F, Adamson GD, Martindale JP, Cuschieri A. Competence assessment of laparoscopic operative and cognitive skills: Objective Structured Clinical Examination (OSCE) or Observational Clinical Human Reliability Assessment (OCHRA). *World J Surg*. 2006 Apr;30(4):527-34.

90. Pandey V, Wolfe JHN, Moorthy K, Munz Y, Jackson MJ, Darzi AW. Technical skills continue to improve beyond surgical training. *Journal of Vascular Surgery*. 2006;43(3):539-45.

91. Moorthy K, Munz Y, Adams S, Pandey V, Darzi A. A human factors analysis of technical and team skills among surgical trainees during procedural simulations in a simulated operating theatre. *Ann Surg.* 2005 Nov;242(5):631-9.
92. Hseino H, Nugent E, Moneley D, Hill ADK, Tierney S. Skills transfer after proficiency-based bench model simulation training in saphenofemoral junction dissection. *JASGBI.* 2012;36:54-6.
93. Naeem N. Validity, reliability, feasibility, acceptability and educational impact of direct observation of procedural skills (DOPS). *J Coll Physicians Surg Pak.* 2013 Jan;23(1):77-82.
94. Wilkinson JR, Crossley JGM, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. *Medical Education.* 2008;42(4):364-73.
95. Morris MC, Gallagher TK, Ridgway PF. Tools used to assess medical students competence in procedural skills at the end of a primary medical degree: a systematic review. *Med Educ Online.* 2012;17.
96. Sarker SK, Maciocco M, Zaman A, Kumar I. Operative performance in laparoscopic cholecystectomy using the Procedural-Based Assessment tool. *American Journal of Surgery.* 2010;200(3):334-40.
97. Howells NR, Carr AJ, Price A, Rees JL. Transferring simulated arthroscopic skills to the operating theatre: a randomised blinded study. *Journal of Bone & Joint Surgery British.* 2008;90(4):494-9.
98. Miller A, Archer J. Impact of workplace based assessment on doctors' education and performance: a systematic review. *bmj.*341.
99. Faulkner H, Regehr G, Martin J, Reznick R. Validation of an objective structured assessment of technical skill for surgical residents. *Acad Med.* 1996 Dec;71(12):1363-5.
100. Van Hove PD, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS, Dankelman J. Objective assessment of technical surgical skills. *British Journal of Surgery.* 2010;97(7):972-87.
101. Martin J.A RG, Reznick R, Macrae H, Murnaghan, Hutchison C, Brown M. objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery.* 1997;84:273-8.
102. Reznick R, Regehr G, MacRae H, Martin J, McCulloch W. Testing technical skill via an innovative "bench station" examination. *Am J Surg.* 1997 Mar;173(3):226-30.
103. Goff BA, Nielsen PE, Lentz GM, Chow GE, Chalmers RW, Fenner D, et al. Surgical skills assessment: a blinded examination of obstetrics and gynecology residents. *Am J Obstet Gynecol.* 2002 Apr;186(4):613-7.
104. Datta V, Bann S, Mandalia M, Darzi A. The surgical efficiency score: a feasible, reliable, and valid method of skills assessment. *American Journal of Surgery.* 2006;192(3):372-8.
105. Siddiqui NY, Stepp KJ, Lasch SJ, Mangel JM, Wu JM. Objective structured assessment of technical skills for repair of fourth-degree perineal lacerations. *Am J Obstet Gynecol.* 2008 Dec;199(6):676 e1-6.
106. Beard J, Choksy S, Khan S. Assessment of operative competence during carotid endarterectomy. *British Journal of Surgery.* 2007;94:726.
107. Datta V, Chang A, Mackay S, Darzi A. The relationship between motion analysis and surgical technical assessments. *American Journal of Surgery.* 2002;184(1):70-3.

108. Goff BA, VanBlaricom A, Mandel L, Chinn M, Nielsen P. Comparison of objective, structured assessment of technical skills with a virtual reality hysteroscopy trainer and standard latex hysteroscopy model. *J Reprod Med*. 2007 May;52(5):407-12.
109. Dath D, Regehr G, Birch D, Schlachta C, Poulin E, Mamazza J, et al. Toward reliable operative assessment: the reliability and feasibility of videotaped assessment of laparoscopic technical skills. *Surg Endosc*. 2004 Dec;18(12):1800-4.
110. Chaer RA, Derubertis BG, Lin SC, Bush HL, Karwowski JK, Birk D, et al. Simulation improves resident performance in catheter-based intervention: results of a randomized, controlled study. *Ann Surg*. 2006 Sep;244(3):343-52.
111. Berry M, Lystig T, Beard J, Klingestierna H, Reznick R, Lönn L. Porcine Transfer Study: Virtual Reality Simulator Training Compared with Porcine Training in Endovascular Novices. *Cardiovasc Intervent Radiol*. 2007 2007/06/01;30(3):455-61.
112. Hislop SJ, Hsu JH, Narins CR, Gillespie BT, Jain RA, Schippert DW, et al. Simulator assessment of innate endovascular aptitude versus empirically correct performance. *Journal of vascular surgery : official publication, the Society for Vascular Surgery [and] International Society for Cardiovascular Surgery, North American Chapter*. 2006;43(1):47-55.
113. Brydges R, Sidhu R, Park J, Dubrowski A. Construct validity of computer-assisted assessment: quantification of movement processes during a vascular anastomosis on a live porcine model. *Am J Surg*. 2007 Apr;193(4):523-9.
114. : Royal College of Obstetrics and Gynaecologists; [cited 2013 9th May ]; Available from: <http://www.rcog.org.uk/files/rcog-corp/uploaded-files/Ed-Core-OSATS.pdf>.
115. Duran CA, Shames M, Bismuth J, Lee JT. Validated assessment tool paves the way for standardized evaluation of trainees on anastomotic models. *Ann Vasc Surg*. 2014 Jan;28(1):115-21.
116. Price J, Naik V, Boodhwani M, Brandys T, Hendry P, Lam B-K. A randomized evaluation of simulation training on performance of vascular anastomosis on a high-fidelity in vivo model: The role of deliberate practice. *The Journal of Thoracic and Cardiovascular Surgery*. 2011;142(3):496-503.
117. Beard J, Bussey M. THE ISCP LEARNING PORTFOLIO. *Bulletin of the Royal College of Surgeons of England*. 2007;89(6):200-2.
118. Pereira EA, Dean BJ. British surgeons' experiences of mandatory online workplace-based assessment. *Journal of the Royal Society of Medicine*. 2009;102(7):287-93.
119. James K, Cross K, Lucarotti ME, Fowler AL, Cook TA. Undertaking procedure-based assessment is feasible in clinical practice. *Annals of the Royal College of Surgeons of England*. 2009;91(2):110-2.
120. Thornton M, Donlon M, Beard JD. The operative skills of higher surgical trainees: measuring competence achieved rather than experience undertaken. *Bulletin of the Royal College of Surgeons of England*. 2003;85(6):190-218.
121. Osborne AJ, Hawkins SC, Pournaras DJ, Chandratilake M, Welbourn R. An evaluation of operative self-assessment by UK postgraduate trainees. *Medical Teacher*. 2014;36(1):32-7.
122. Duschek N, Assadian A, Lamont PM, Klemm K, Schmidli J, Mendel H, et al. Simulator training on pulsatile vascular models significantly improves surgical skills and the quality of carotid patch plasty. *Journal of vascular surgery : official*

- publication, the Society for Vascular Surgery [and] International Society for Cardiovascular Surgery, North American Chapter. 2013;57(4):1148-54.
123. Sutherland LM, Middleton PF, Anthony A, Hamdorf J, Cregan P, Scott D, et al. Surgical simulation: a systematic review. *Ann Surg*. 2006 Mar;243(3):291-300.
124. Cameron JL. William Stewart Halsted. Our surgical heritage. *Ann Surg*. 1997 May;225(5):445-58.
125. Aggarwal R, Mytton OT, Derbrew M, Hananel D, Heydenburg M, Issenberg B, et al. Training and simulation for patient safety. *Quality and Safety in Health Care*. 2010 August 1, 2010;19(Suppl 2):i34-i43.
126. Peters S, Bird L, Ashraf H, Ahmed S, McNamee P, Ng C, et al. Medical undergraduates' use of behaviour change talk: the example of facilitating weight management. *BMC Medical Education*. 2013;13(1):7.
127. Fonseca AL, Evans LV, Gusberg RJ. Open Surgical Simulation in Residency Training: A Review of Its Status and a Case for Its Incorporation. *Journal of Surgical Education*. 2013;70(1):129-37.
128. Black SA, Harrison RH, Horrocks EJ, Pandey VA, Wolfe JHN. Competence assessment of senior vascular trainees using a carotid endarterectomy bench model. *British Journal of Surgery*. 2007;94(10):1226-31.
129. Scott DJ, Bergen PC, Rege RV, Laycock R, Tesfay ST, Valentine RJ, et al. Laparoscopic training on bench models: better and more cost effective than operating room experience? *J Am Coll Surg*. 2000 Sep;191(3):272-83.
130. Shah J, Mackay S, Vale J, Darzi A. Simulation in urology--a role for virtual reality? *BJU Int*. 2001 Nov;88(7):661-5.
131. Burden C, Oestergaard J, Larsen CR. Integration of laparoscopic virtual-reality simulation into gynaecology training. *BJOG*. 2011 Nov;118 Suppl 3:5-10.
132. Symbionix. Symbionix (TM) To Advance Clinical Performance. [cited 2103 18th May]; Website]. Available from: <http://symbionix.com/simulators/>.
133. Van Herzeele I, Aggarwal R. Virtual Reality Simulation in the Endovascular Field. European Virtual Reality Endovascular Research Team (EVEREST) US Cardiology. 2008;5(1):41-5.
134. Perkins GD. Simulation in resuscitation training. *Resuscitation*. 2007;73(2):202-11.
135. Walsh CM, Sherlock ME, Ling SC, Carnahan H. Virtual reality simulation training for health professions trainees in gastrointestinal endoscopy. *Cochrane Database Syst Rev*. 2012;6:CD008237.
136. Webster PC. Richard Reznick: leading innovator of surgical education. *The Lancet*. 2011;378(9785):21.
137. Sonnadara RR, Garbedian S, Safir O, Nousiainen M, Alman B, Ferguson P, et al. Orthopaedic Boot Camp II: examining the retention rates of an intensive surgical skills course. *Surgery*. 2012 Jun;151(6):803-7.
138. Sonnadara RR, Garbedian S, Safir O, Mui C, Mironova P, Nousiainen M, et al. Toronto orthopaedic boot camp III: examining the efficacy of student-regulated learning during an intensive, laboratory-based surgical skills course. *Surgery*. 2013 Jul;154(1):29-33.
139. Sonnadara RR, Van Vliet A, Safir O, Alman B, Ferguson P, Kraemer W, et al. Orthopedic boot camp: examining the effectiveness of an intensive surgical skills course. *Surgery*. 2011 Jun;149(6):745-9.
140. Tsang JS, Naughton PA, Leong S, Hill ADK, Kelly CJ, Leahy AL. Virtual reality simulation in endovascular surgical training. *The Surgeon*. 2008;6(4):214-20.

141. Lamont PM, Scott DJA. The impact of shortened training times on the discipline of vascular surgery in the United Kingdom. *The American Journal of Surgery*. 2005;190(2):269-72.
142. Bevan PG. The first anastomosis workshop, March 1981. *Ann R Coll Surg Engl*. 1997 Jul;79(4 Suppl):168-9.
143. Datta V, Bann S, Aggarwal R, Mandalia M, Hance J, Darzi A. Technical skills examination for general surgical trainees. *British Journal of Surgery*. 2006;93(9):1139-46.
144. Black SA, Nestel DF, Kneebone RL, Wolfe JHN. Assessment of surgical competence at carotid endarterectomy under local anaesthesia in a simulated operating theatre. *British Journal of Surgery*. 2010;97(4):511-6.
145. Lentz GM, Mandel LS, Lee D, Gardella C, Melville J, Goff BA. Testing surgical skills of obstetric and gynecologic residents in a bench laboratory setting: validity and reliability. *Am J Obstet Gynecol*. 2001 Jun;184(7):1462-8; discussion 8-70.
146. Shamim Khan M, Ahmed K, Gavazzi A, Gohil R, Thomas L, Poulsen J, et al. Development and implementation of centralized simulation training: evaluation of feasibility, acceptability and construct validity. *BJU International*. 2013;111(3):518-23.
147. Leong JH, Leff DR, Das A, Aggarwal R, Reilly P, Atkinson HDE, et al. Validation of orthopaedic bench models for trauma surgery. *Journal of Bone & Joint Surgery, British Volume*. 2008 July 1, 2008;90-B(7):958-65.
148. Goff B, Mandel L, Lentz G, Vanblaricom A, Oelschlager AM, Lee D, et al. Assessment of resident surgical skills: is testing feasible? *Am J Obstet Gynecol*. 2005 Apr;192(4):1331-8; discussion 8-40.
149. Anastakis DJ, Regehr G, Reznick RK, Cusimano M, Murnaghan J, Brown M, et al. Assessment of technical skills transfer from the bench training model to the human model. *Am J Surg*. 1999 Feb;177(2):167-70.
150. Hamilton EC, Scott DJ, Kapoor A, Nwariaku F, Bergen PC, Rege RV, et al. Improving operative performance using a laparoscopic hernia simulator. *Am J Surg*. 2001 Dec;182(6):725-8.
151. Matsumoto ED, Hamstra SJ, Radomski SB, Cusimano MD. The effect of bench model fidelity on endourological skills: a randomized controlled study. *J Urol*. 2002 Mar;167(3):1243-7.
152. Youngblood PL, Srivastava S, Curet M, Heinrichs WL, Dev P, Wren SM. Comparison of training on two laparoscopic simulators and assessment of skills transfer to surgical performance. *J Am Coll Surg*. 2005 Apr;200(4):546-51.
153. Keyser EJ, Derossis AM, Antoniuk M, Sigman HH, Fried GM. A simplified simulator for the training and evaluation of laparoscopic skills. *Surg Endosc*. 2000 Feb;14(2):149-53.
154. Jordan JA, Gallagher AG, McGuigan J, McClure N. Randomly alternating image presentation during laparoscopic training leads to faster automation to the "fulcrum effect". *Endoscopy*. 2000 Apr;32(4):317-21.
155. Harold KL, Matthews BD, Backus CL, Pratt BL, Heniford BT. Prospective randomized evaluation of surgical resident proficiency with laparoscopic suturing after course instruction. *Surg Endosc*. 2002 Dec;16(12):1729-31.
156. Risucci D, Cohen JA, Garbus JE, Goldstein M, Cohen MG. The effects of practice and instruction on speed and accuracy during resident acquisition of simulated laparoscopic skills. *Curr Surg*. 2001 Mar;58(2):230-5.

157. Pandey VA, Black SA, Lazaris AM, Allenberg JR, Eckstein H. H, Hagmuller, et al. Do Vascular Workshops Improve the Technical Skills of Vascular Surgical Trainees? *European Journal of Vascular and Endovascular surgery*. 2005;30:441-7.
158. Pandey VA, Wolfe JHN, Liapis CD, Bergqvist D. The examination assessment of technical competence in vascular surgery. *British Journal of Surgery*. 2006;93(9):1132-8.
159. Bann S, Kwok KF, Lo CY, Darzi A, Wong J. Objective assessment of technical skills of surgical trainees in Hong Kong. *Br J Surg*. 2003 Oct;90(10):1294-9.
160. Sidhu RS, Park J, Brydges R, MacRae HM, Dubrowski A. Laboratory-based vascular anastomosis training: a randomized controlled trial evaluating the effects of bench model fidelity and level of training on skill acquisition. *J Vasc Surg*. 2007 Feb;45(2):343-9.
161. Robinson WP, 3rd, Schanzer A, Cutler BS, Baril DT, Larkin AC, Eslami MH, et al. A randomized comparison of a 3-week and 6-week vascular surgery simulation course on junior surgical residents' performance of an end-to-side anastomosis. *J Vasc Surg*. 2012 Dec;56(6):1771-80; discussion 80-1.
162. Robinson WP, Baril DT, Taha O, Schanzer A, Larkin AC, Bismuth J, et al. Simulation-based training to teach open abdominal aortic aneurysm repair to surgical residents requires dedicated faculty instruction. *J Vasc Surg*. 2013 Jul;58(1):247-53 e1-2.
163. Dawson SL, Cotin S, Meglan D, Shaffer DW, Ferrell MA. Designing a computer-based simulator for interventional cardiology training. *Catheterization and Cardiovascular Interventions*. 2000;51(4):522-7.
164. Berry M, Reznick R, Lystig T, Lonn L. The use of virtual reality for training in carotid artery stenting: a construct validation study. *Acta Radiol*. 2008 Sep;49(7):801-5.
165. Dawson DL, Meyer J, Lee ES, Pevac WC. Training with simulation improves residents' endovascular procedure skills. *J Vasc Surg*. 2007 Jan;45(1):149-54.
166. Dayal R, Faries PL, Lin SC, Bernheim J, Hollenbeck S, DeRubertis B, et al. Computer simulation as a component of catheter-based training. *J Vasc Surg*. 2004 Dec;40(6):1112-7.
167. Hsu JH, Younan D, Pandalai S, Gillespie BT, Jain RA, Schippert DW, et al. Use of computer simulation for determining endovascular skill levels in a carotid stenting model. *Journal of Vascular Surgery*. 2004;40(6):1118-25.
168. Berry M, Lystig T, Reznick R, Lonn L. Assessment of a Virtual Interventional Simulator Trainer. *Journal of Endovascular Therapy*. 2006 2006/04/01;13(2):237-43.
169. Neequaye SK, Aggarwal R, Brightwell R, Van Herzelee I, Darzi A, Cheshire NJW. Identification of Skills Common to Renal and Iliac Endovascular Procedures Performed on a Virtual Reality Simulator. *European journal of vascular and endovascular surgery : the official journal of the European Society for Vascular Surgery*. 2007;33(5):525-32.
170. Lee JT, Peruzzaro A, Krummel T, Dalman RL. SS22. See One, Sim One, Do One, Teach One: Results of a Prospective Randomized Trial of Endovascular Skills Training for Surgical Residents. *Journal of Vascular Surgery*. 55(6):27S.
171. Patel AD, Gallagher AG, Nicholson WJ, Cates CU. Learning Curves and Reliability Measures for Virtual Reality Simulation in the Performance Assessment

- of Carotid Angiography. *Journal of the American College of Cardiology*. 2006;47(9):1796-802.
172. Aggarwal R, Black SA, Hance JR, Darzi A, Cheshire NJW. Virtual Reality Simulation Training can Improve Inexperienced Surgeons' Endovascular Skills. *European Journal of Vascular and Endovascular Surgery*. 2006;31(6):588-93.
173. Van Herzelee I, Aggarwal R, Choong A, Brightwell R, Vermassen FE, Cheshire NJ. Virtual reality simulation objectively differentiates level of carotid stent experience in experienced interventionalists. *Journal of Vascular Surgery*. 2007;46(5):855-63.
174. Van Herzelee I, Aggarwal R, Neequaye S, Hamady M, Cleveland T, Darzi A, et al. Experienced endovascular interventionalists objectively improve their skills by attending carotid artery stent training courses. *Eur J Vasc Endovasc Surg*. 2008 May;35(5):541-50.
175. Intensive endovascular simulation training improves performance. *Vascular NEWS international*. 2012 September:1-2.
176. Grantcharov TP, Schulze S, Kristiansen VB. The impact of objective assessment and constructive feedback on improvement of laparoscopic performance in the operating room. *Surg Endosc*. 2007 Dec;21(12):2240-3.
177. Hu Y-Y, Peyre SE, Arriaga AF, Osteen RT, Corso KA, Weiser TG, et al. Postgame Analysis: Using Video-Based Coaching for Continuous Professional Development. *Journal of the American College of Surgeons*. 2012;214(1):115-24.
178. Sigounas VY, Callas PW, Nicholas C, Adams JE, Bertges DJ, Stanley AC, et al. Evaluation of simulation-based training model on vascular anastomotic skills for surgical residents. *Simul Healthc*. 2012 Dec;7(6):334-8.
179. Sedgwick P. Parametric v non-parametric statistical tests. *bmj*. 2012 2012-03-14 16:03:33;344.
180. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986 Feb 8;1(8476):307-10.
181. The Human Tissue Act: Chapter 30. The National Archives; 2004.
182. Marx RG, Menezes A, Horovitz L, Jones EC, Warren RF. A comparison of two time intervals for test-retest reliability of health status instruments. *J Clin Epidemiol*. 2003 Aug;56(8):730-5.
183. Slabach E, H, Morrow J, Wache T, D. Questionnaire Measurement of infant and child temperament: Current status and future directions. In: Jan Strelau, Angleitner A, editors. *Explorations in Temperament: International Perspectives on Theory and Measurement*. New York: Springer Science US; 1991. p. 205-24.
184. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. *Acad Emerg Med*. 2008 Nov;15(11):988-94.
185. Gawande AA, Zinner MJ, Studdert DM, Brennan TA. Analysis of errors reported by surgeons at three teaching hospitals. *Surgery*. 2003;133(6):614-21.
186. Yule S, Flin R, Paterson-Brown S, Maran N. Non-technical skills for surgeons in the operating room: A review of the literature. *Surgery*. 139(2):140-9.
187. Mandel LS, Goff BA, Lentz GM. Self-assessment of resident surgical skills: is it feasible? *Am J Obstet Gynecol*. 2005 Nov;193(5):1817-22.
188. Moorthy K, Munz Y, Adams S, Pandey V, Darzi A. Self-assessment of performance among surgical trainees during simulated procedures in a simulated operating theater. *The American Journal of Surgery*. 2006;192(1):114-8.

189. Lipsett Pa HIDS. Resident self-other assessor agreement: Influence of assessor, competency, and performance level. *Archives of Surgery*. 2011;146(8):901-6.
190. Aggarwal R, Grantcharov T, Moorthy K, Milland T, Papasavas P, Dosis A, et al. An evaluation of the feasibility, validity, and reliability of laparoscopic skills assessment in the operating room. *Ann Surg*. 2007 Jun;245(6):992-9.
191. Vassiliou MC, Ghitulescu GA, Feldman LS, Stanbridge D, Leffondre K, Sigman HH, et al. The MISTELS program to measure technical skill in laparoscopic surgery : evidence for reliability. *Surg Endosc*. 2006 May;20(5):744-7.
192. Beard JD, Jolly BC, Newble DI, Thomas WEG, Donnelly J, Southgate LJ. Assessing the technical skills of surgical trainees. *British Journal of Surgery*. 2005;92(6):778-82.
193. Khan MW, Lin D, Marlow N, Aintree M, Babidge W, Field J, et al. Laparoscopic Skills Maintenance: A Randomized Trial of Virtual Reality and Box Trainer Simulators. *Journal of Surgical Education*. 2014;71(1):79-84.
194. Collins AM, Ridgway PF, Hassan MSU, Chou CWK, Hill AD, Kneafsey B. Surgical instruction for general practitioners: How, who and how often? *Journal of plastic, reconstructive & aesthetic surgery : JPRAS*. 2010;63(7):1156-62.
195. Weisz G, Smilowitz NR, Parise H, Devaud J, Moussa I, Ramee S, et al. Objective Simulator-Based Evaluation of Carotid Artery Stenting Proficiency (from Assessment of Operator Performance by the Carotid Stenting Simulator Study [ASSESS]). *The American journal of cardiology*. 2013;112(2):299-306.
196. Spiotta AM, Rasmussen PA, Masaryk TJ, Benzel EC, Schlenk R. Simulated diagnostic cerebral angiography in neurosurgical training: a pilot program. *J Neurointerv Surg*. 2013 Jul;5(4):376-81.
197. Moulton CA, Dubrowski A, Macrae H, Graham B, Grober E, Reznick R. Teaching surgical skills: what kind of practice makes perfect?: a randomized, controlled trial. *Ann Surg*. 2006 Sep;244(3):400-9.
198. Lammers RL, Davenport M, Korley F, Griswold-Theodorson S, Fitch MT, Narang AT, et al. Teaching and Assessing Procedural Skills Using Simulation: Metrics and Methodology. *Academic Emergency Medicine*. 2008;15(11):1079-87.
199. Strandbygaard J, Bjerrum F, Maagaard M, Winkel P, Larsen CR, Ringsted C, et al. Instructor Feedback Versus No Instructor Feedback on Performance in a Laparoscopic Virtual Reality Simulator: A Randomized Trial. *Annals Of Surgery*. 2013;257(5):839-44 10.1097/SLA.0b013e31827eee6e.
200. Satpathy A KR. Workplaced Based Surgical Assessments: assessment for learning or assessment of learning? Surgeons' perceptions of learning benefits of assessment in surgical training. . *British Journal of Surgery*. [Published Abstracts]. 2010;97((S2)):115.
201. Beard PJD. PBA Generic assessment template (revised v2).pdf. In: Green MLA, editor. email ed2015.
202. Learning and assessment in the clinical environment: the way forward: GMC2011.
203. JCST Strategy 2013-2018 JCST- Joint Committee on Surgical Training2013.

## Appendix 1-ICEPS

Imperial  
College  
Evaluation of  
Procedure-specific  
Skill

### Saphenofemoral Junction Ligation

Candidate no:  
Assessor:  
Date:

Please circle the candidate's performance on the following scale:

	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<b>Incision</b>	<b>1</b> Does not use surface landmarks. Inappropriate placement of incision. Poor handling of scalpel	<b>2</b>	<b>3</b> Appropriate incision in terms of location and size. Looked at ease with scalpel	<b>4</b>	<b>5</b> Uses surface landmarks to make an appropriately located and sized incision. Handled scalpel expertly
<b>Dissection</b>	<b>1</b> Appeared unsure and excessively hesitant whilst dissecting. Caused trauma to tissues. Did not dissect into the correct anatomical plane.	<b>2</b>	<b>3</b> Controlled and safe dissection into correct anatomical plane. Caused minimum trauma of tissues. Used instrument satisfactorily while dissecting	<b>4</b>	<b>5</b> Superior and atraumatic dissection into the correct anatomical plane. Confident handling of instruments whilst dissecting.
<b>Retraction</b>	<b>1</b> Clumsy use of retractors. Did not allow visualization of the important structures making frequent changes to retractor setting.	<b>2</b>	<b>3</b> Good use of retraction allowing visualization of major structures. Had to change retractor position to visualize other structures	<b>4</b>	<b>5</b> Excellent use of retractors. Allowed good visualization of all necessary structures. Atraumatic
<b>Tributaries</b>	<b>1</b> Could not or did not try to identify any tributaries	<b>2</b>	<b>3</b> Identified all known tributaries. Did not seek other vessels	<b>4</b>	<b>5</b> Identified all known tributaries. Sought other possible tributaries.
<b>Haemostasis</b>	<b>1</b> Poor quality of knot tying. Knots frequently slipped or was excessively traumatic to vessels	<b>2</b>	<b>3</b> Competent knot tying. Minimal trauma to vessels. Minimal blood loss.	<b>4</b>	<b>5</b> Superior knot tying. Atraumatic. No knot slippage
<b>SFJ Clearance</b>	<b>1</b> Did not identify the Saphenofemoral junction or excessively traumatic dissection around that vessel	<b>2</b>	<b>3</b> Identified the Saphenofemoral junction. Safely dissected tissues away from vessel. Reasonable clearance of vessel. Minimal trauma.	<b>4</b>	<b>5</b> Identified the Saphenofemoral junction. Expert dissection of tissues off the vessels. Atraumatic. Cleared well proximally and distally.
<b>SFJ Ligation</b>	<b>1</b> Did not ligate the SFJ or ligated CFV or caused excessive encroachment onto CFV after SFJ Ligation	<b>2</b>	<b>3</b> Good knot tying whilst ligating the SFJ. Minimal encroachment onto CFV following SFJ ligation.	<b>4</b>	<b>5</b> Excellent safe and secure ligation of the SFJ. Flush ligation with no encroachment onto CFV

**Total score:**

## Appendix 2- DOPS

Trainee		Assessor	
Name:		Name:	
GMC/GDC/IMC number:		GMC/GDC/IMC number:	
Speciality:		Position:	
Hospital/Organisation:		Institutional e-mail:	
Training level:		Training: No <input type="checkbox"/> Written <input type="checkbox"/> Web/CD <input type="checkbox"/> Workshop <input type="checkbox"/>	
Name of procedure:			
Difficulty of procedure: Easier than usual <input type="checkbox"/> Average difficulty <input type="checkbox"/> More difficult than usual <input type="checkbox"/>			
Performed in a simulated setting <input type="checkbox"/>		Number of times procedure previously performed:	
Rate the domains using the following scale: N = Not observed or not appropriate, D = Development required, S = Satisfactory standard for completion of CCST (no prompting or intervention required)			
Domain	Rating ND/S	Comments	
1. Describes indications, anatomy, procedure and complications to assessor			
2. Obtains consent, after explaining procedure and possible complications to patient			
3. Prepares for procedure according to an agreed protocol			
4. Administers effective analgesia or safe sedation (if no anaesthetist)			
5. Demonstrates good asepsis and safe use of instruments and sharps			
6. Performs the technical aspects in line with the guidance notes			
7. Deals with any unexpected event or seeks help when appropriate			
8. Completes required documentation (written or dictated)			
9. Communicates clearly with patient & staff throughout the procedure			
10. Demonstrates professional behaviour throughout the procedure			
<b>FEEDBACK:</b> Verbal feedback is a mandatory component of this assessment. Please use this space to record areas of strength and suggestions for development which were highlighted during discussion with the trainee:			
<b>GLOBAL SUMMARY</b> After summarising the discussion with the trainee in the box above, please complete the level at which the procedure was performed on this occasion			Tick
Level 0	Insufficient evidence observed to support a summary judgement		
Level 1	Unable to perform the procedure under supervision		
Level 2	Able to perform the procedure under supervision		
Level 3	Able to perform the procedure with minimum supervision (needed occasional help)		
Level 4	Competent to perform the procedure unsupervised (and could deal with any complications that arose)		
Time taken for observation (mins):		Time taken for feedback (mins):	
Date:	Trainee's signature:	Assessor's signature:	
Trainee satisfaction with CbD	Assessor satisfaction with CbD		
1 <input type="checkbox"/>	1 <input type="checkbox"/>		
2 <input type="checkbox"/>	2 <input type="checkbox"/>		
3 <input type="checkbox"/>	3 <input type="checkbox"/>		
4 <input type="checkbox"/>	4 <input type="checkbox"/>		
5 <input type="checkbox"/>	5 <input type="checkbox"/>		
6 <input type="checkbox"/>	6 <input type="checkbox"/>		
7 <input type="checkbox"/>	7 <input type="checkbox"/>		
8 <input type="checkbox"/>	8 <input type="checkbox"/>		
9 <input type="checkbox"/>	9 <input type="checkbox"/>		
10 <input type="checkbox"/>	10 <input type="checkbox"/>		

Updated 1.12

## Appendix 3-An example of the original OSATS tool and generic OSATS

Example taken from Martin J. A. et al Objective Structured Assessment of Surgical Skill (OSATS) for surgical residents(101)

OBJECTIVE STRUCTURED ASSESSMENT OF TECHNICAL SKILL 27

RESIDENT  
STICKER

**STATION 4**

**CONTROL OF HAEMORRHAGE**

**INSTRUCTIONS TO CANDIDATES**

You have just identified a stab wound to the inferior vena cava. Control the haemorrhage and repair the vessel.

Start Time: .....

**CHECKLIST**

ITEM	Not Done/ Done Incorrectly	Done Correctly
<b>CONTROL OF HEMORRHAGE</b>		
1. Applies pressure to stop bleeding first	0	1
2. Asks assistant to suction field	0	1
3. Inspects injury by carefully releasing the IVC	0	1
4. Ensures all equipment needed for repair is at hand before starting	0	1
5. Control of bleeding point (use deBakey forceps /Satinsky clamp or prox/distal pressure)	0	1
<b>REPAIR</b>		
6. Select appropriate suture (4.0/5.0/6.0 polypropylene)	0	1
7. Select appropriate needle driver (vascular)	0	1
8. Select appropriate forceps (de Bakey)	0	1
9. Needle loaded 1/2-2/3 from tip 90% of time	0	1

1 of a task-specific checklist used in the Objective Structured Assessment of Technical Skill. IVC, inferior vena cava

Please rate the candidate's performance on the following scale:

	1	2	3	4	5
<b>Respect for tissue</b>	Frequently used unnecessary force on tissue or caused damage by inappropriate use of instruments.		Careful handling of tissue but occasionally caused inadvertent damage.		Consistently handled tissues appropriately with minimal damage.
<b>Time and motion</b>	Many unnecessary moves.		Efficient time/motion but some unnecessary moves.		Economy of movement and maximum efficiency.
<b>Instrument handling</b>	Repeatedly makes tentative or awkward moves with instruments.		Competent use of instruments although occasionally appeared stiff or awkward.		Fluid moves with instruments and no awkwardness.
<b>Knowledge of instruments</b>	Frequently asked for the wrong instrument or used an inappropriate instrument.		Knew the names of most instruments and used appropriate instrument for the task.		Obviously familiar with the instruments required and their names.
<b>Use of assistants</b>	Consistently placed assistants poorly or failed to use assistants.		Good use of assistants most of the time.		Strategically used assistant to the best advantage at all times.
<b>Flow of operation and forward planning</b>	Frequently stopped operating or needed to discuss next move.		Demonstrated ability for forward planning with steady progression of operative procedure.		Obviously planned course of operation with effortless flow from one move to the next.
<b>Knowledge of specific procedure</b>	Deficient knowledge. Needed specific instruction at most operative steps.		Knew all important aspects of the operation.		Demonstrated familiarity with all aspects of the operation.

Overall, on this task, should this candidate:  Pass  Fail ?

Fig. 2 Detailed global 5-point rating scale and pass/failure score for Objective Structured Assessment of Technical Skill

### Generic OSATS used to assess concurrent validity

Respect for Tissues					Score
1	2	3	4	5	
Handles tissue poorly causing frequent damage		Handles tissue carefully with occasional damage		Handles tissue well with minimal damage	
Efficiency of Motion					
1	2	3	4	5	
Inefficient movements and many unnecessary moves		Efficient movements with some unnecessary moves		Highly efficient movement with very few unnecessary moves	
Instrument Handling					
1	2	3	4	5	
Repeatedly makes tentative or awkward moves with instrument		Competent use of instruments although occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness	
Use of assistants					
1	2	3	4	5	
Consistently placed assistants poorly or failed to use assistants		Good use of assistants most of the time		Strategically used assistant to the best advantage at all times	
Knowledge of instruments					
1	2	3	4	5	
Frequently asked for the wrong instrument or used inappropriate instrument		Knew the name of most of the instruments and used appropriate instrument for task		Obviously familiar with the instruments required and their names	
Knowledge of specific procedure					
1	2	3	4	5	
Deficient knowledge, needed specific instruction at most operative steps		Knew all important aspects of the operation		Demonstrated familiarity with all aspect of the operation	
Flow of operation					
1	2	3	4	5	
frequent pauses in operation with uncertainty about next step		some pauses in operation with little uncertainty about next step		well planned operation with minimal uncertainty about next step	

## Appendix 4- PBA

Vascular Surgery PBA: Infra-inguinal Bypass		
Trainee:	Assessor:	Date:
Assessor's Position*:	Email (institutional):	GMC No:
Duration of procedure (mins):	Duration of assessment period (mins):	Hospital:
Operation more difficult than usual? Yes / No (if yes, state reason)	; Tick this box if this PBA was performed in a <b>Simulated setting</b>	
Complexity (tick which applies, if any)	<input type="checkbox"/> Basic (femoropopliteal bypass - AK) <input type="checkbox"/> Intermediate (femoropopliteal bypass - BK) <input type="checkbox"/> Advanced (femorofemoral bypass)	

\* Assessors are normally consultants (senior trainees may be assessors depending upon their training level and the complexity of the procedure)

**IMPORTANT:** The trainee should explain what he/she intends to do throughout the procedure. The Assessor should provide verbal prompts if required, and intervene if patient safety is at risk.

### Rating:

**N** = Not observed or not appropriate

**D** = Development required

**S** = Satisfactory standard for CCT (no prompting or intervention required)

Competencies and Definitions		Rating NUS	Comments
<b>I. Consent</b>			
C1	Demonstrates sound knowledge of indicators and contraindications including alternatives to surgery		
C2	Demonstrates awareness of sequelae of operative or non operative management.		
C3	Demonstrates sound knowledge of complications of surgery		
C4	Explains the procedure to the patient / relatives / carers and checks understanding		
C5	Explains likely outcome and time to recovery and checks understanding		
<b>II. Pre operation planning</b>			
PL1	Demonstrates recognition of anatomical and pathological abnormalities (and relevant co-morbidities) and selects appropriate operative strategies / techniques to deal with these		
PL2	Demonstrates ability to make reasoned choice of appropriate equipment, materials or devices (if any) taking into account appropriate investigations e.g. x-rays		
PL3	Checks materials, equipment and device requirements with operating room staff		
PL4	Ensures the operation site is marked where applicable (including vein if used)		
PL5	Checks patient records, and ensures that all investigations including images are available		
<b>III. Pre operative preparation</b>			
PR1	Checks in theatre that consent has been obtained		
PR2	Gives effective briefing to theatre team		
PR3	Ensures proper and safe positioning of the patient on the operating table		
PR4	Demonstrates careful skin preparation		
PR5	Demonstrates careful draping of the patient's operative field		
PR6	Ensures general equipment and materials are deployed safely (e.g. catheter, clamping)		
PR7	Ensures appropriate drugs administered (e.g. prophylactic antibiotics)		
PR8	Arranges for and deploys specialist equipment (e.g. image intensifiers) effectively		
<b>IV. Exposure and closing</b>			
E1	Demonstrates knowledge of optimum skin incision / portal / access		

E2	Achieves an adequate exposure through purposeful dissection in correct tissue planes and identifies all structures correctly	
E3	Completes a sound wound repair where appropriate	
E4	Protects the wound with dressings, splints and drains where appropriate	
<b>V. Intra operative technique: global (G) and task-specific items (T)</b>		
IT1(G)	Follows an agreed, logical sequence or protocol to the procedure	
IT2(G)	Consistently handles tissue well with minimal damage	
IT3(G)	Controls bleeding promptly by an appropriate method	
IT4(G)	Demonstrates a sound technique of knots and sutures/stapes	
IT5(G)	Uses instruments appropriately and safely	
IT6(G)	Proceeds at appropriate pace with economy of movement	
IT7(G)	Anticipates and responds appropriately to variation e.g. anatomy	
IT8(G)	Deals calmly and effectively with unexpected events/complications	
IT9(G)	Uses assistant(s) to the best advantage at all times	
IT10(G)	Communicates clearly and consistently with the scrub team	
IT11(G)	Communicates clearly and consistently with the anaesthetist (e.g. Heparin)	
*	Proximal Anastomosis	
IT12(T)	Displays and controls inflow vessel and confirms inflow site	
IT13(T)	Makes suitable arteriotomy at selected inflow site	
IT14(T)	Confirms good inflow and performs endarterectomy or extends arteriotomy if required	
IT15(T)	Performs anastomosis with sound eversion technique	
IT16(T)	Ensures there is no narrowing or distortion of vessels on completion of anastomosis	
*	LSV Harvest (can apply to harvest of other veins and/or tunnelling)	
IT17(T)	Dissects proximal end of LSV and detaches it cleanly from femoral vein	
IT18(T)	Closes femoral vein without bleeding or narrowing	
IT19(T)	Harvests LSV without damage, ligates branches securely and doesn't undermine skin	
IT20(T)	Tunnels graft without damage to surrounding structures and without twisting or kinking	
IT21(T)	Disrupts venous valves without damage to vein wall, tributaries or proximal anastomosis	
*	Distal Anastomosis	
IT22(T)	Displays and controls outflow vessel and confirms outflow site	
IT23(T)	Makes suitable arteriotomy at selected outflow site	
IT24(T)	Confirms good outflow and performs endarterectomy or extends arteriotomy if required	
IT25(T)	Performs anastomosis with sound eversion technique, using loupes if required	
IT26(T)	Ensures there is no narrowing or distortion of vessels on completion of anastomosis	
IT27(T)	Carries out quality check on graft function (e.g. duplex, flow measurement, arteriogram)	
<b>VI. Postoperative management</b>		
PM1	Ensures the patient is transferred safely from the operating table to bed	
PM2	Constructs a clear operation note	
PM3	Records clear and appropriate post operative instructions	
PM4	Deals with specimens. Labels and orientates specimens appropriately - NOT APPLICABLE TO THIS PROCEDURE	

#### Global Summary

Level at which completed elements of the PBA were performed on this occasion		Tick as appropriate
Level 0	Insufficient evidence observed to support a summary judgement	



## Appendix 5 – Modified PBA for Sapheno-femoral junction ligation, proximal aortic anastomosis and superficial femoral artery angioplasty

Vascular Surgery PBA:			
Aortic Aneurysm- proximal anastomosis			
Grade of Trainee:			
Grade of Assessor:			
Number of times procedure seen or performed before	Seen		Done

Rating N= Not observed, D= Development required, S= Satisfactory Standard for CCT

Competencies and Definitions	Rating N/D/S
<b>Exposure and closure</b>	
Demonstrates knowledge of optimum skin incision (position and length)	
Achieves an adequate exposure of aortic neck and left renal vein (without delay if ruptured)	
Completes a sound abdominal wound repair	
Follows an agreed, logical sequence or protocol for the procedure	
Consistently handles tissue well with minimal damage	
Demonstrates a sound technique of knots and sutures/staples	
Uses instruments appropriately and safely	
Proceeds at appropriate pace with economy of movement	
Anticipates and responds appropriately to variation e.g. left renal vein	
Deals calmly and effectively with unexpected events / complications	
Uses assistant(s) to the best advantage at all times	
Selects appropriate level to clamp aorta	
Exposes iliac arteries and identifies appropriate level for clamping	
Clamps iliac arteries without damaging adjacent structures (veins and ureters)	
Clamps aortic neck without damaging adjacent structures (veins and duodenum)	
Opens aortic sac, removes thrombus and controls lumbar back-bleeding without delay	

Prepares neck of aneurysm to receive graft (e.g. T cut) and displays with retractor	
Selects appropriate graft (size and configuration)	
Selects appropriate suture and needle holder	
Shortens body of graft if using a bifurcation graft	
Sutures graft into aorta using appropriately placed sutures without tearing aorta	
Avoids distortion of aorta at "corners"	
Ensures that assistant maintains tension during suturing	
Completes suture line with maintained tension and adequate knot (>6 throws)	
Tests anastomosis, identifies and corrects any defects	
Prepares aortic bifurcation or iliac arteries to receive graft	
Cuts graft to correct length to ensure no redundancy	
Selects appropriate suture and needle holder	
Ensures that assistant maintains tension during suturing	
Completes suture line with maintained tension and adequate knot (>6 throws)	
Tests anastomosis, identifies and corrects any defects	

#### GLOBAL SUMMARY

Level at which completed elements of the PBA were performed on this occasion ( Tick as appropriate)		
Level 0	Insufficient evidence observed to support a summary judgement	
Level 1	Unable to perform the procedure, or part observed, under supervision	
Level 2	Able to perform the procedure, or part observed, under supervision	
Level 3	Able to perform the procedure with minimum supervision (needed occasional help)	
Level 4	Competent to perform the procedure unsupervised (could deal with complications that arose)	

## Vascular Surgery PBA: SFJ ligation

Grade of Trainee:				
Grade of Assessor:				
Number of times procedure seen or performed before <b>BY TRAINEE</b>	Seen		Done	

Rating N= Not observed, D= Development required, S= Satisfactory Standard for CCT

Competencies and Definitions	Rating N/D/S
<b>Exposure and closure</b>	
Demonstrates knowledge of optimum skin incision / portal / access	
Achieves an adequate exposure through purposeful dissection in correct tissue planes and identifies all structures correctly	
Completes a sound wound repair where appropriate	
<b>Global and Task Specific</b>	
Follows an agreed, logical sequence or protocol for the procedure	
Consistently handles tissue well with minimal damage	
Demonstrates a sound technique of knots and sutures/staples	
Uses instruments appropriately and safely	
Proceeds at appropriate pace with economy of movement	
Anticipates and responds appropriately to variation e.g. anatomy	
Deals calmly and effectively with unexpected events/complications	
Uses assistant(s) to the best advantage at all times	
Positions self-retaining retractor to expose the proximal LSV	
Ligates and divides second order tributaries (diathermy permissible for small tributaries)	
Clearly identifies Saphenofemoral junction through cribriform fascia	

Divides long saphenous vein between clips, or ligates, after identification of SFJ	
Inspects SFJ to ensure no missed tributaries	
Flushes ligation or transfixes SFJ	

## GLOBAL SUMMARY

Level at which completed elements of the PBA were performed on this occasion ( Tick as appropriate)		
Level 0	Insufficient evidence observed to support a summary judgement	
Level 1	Unable to perform the procedure, or part observed, under supervision	
Level 2	Able to perform the procedure, or part observed, under supervision	
Level 3	Able to perform the procedure with minimum supervision (needed occasional help)	
Level 4	Competent to perform the procedure unsupervised (could deal with complications that arose)	

## Endovascular PBA: Peripheral artery angioplasty/stenting

<b>Trainee:</b>	<b>Assessor:</b>	<b>Date:</b>
<b>Start time:</b>	<b>End time:</b>	<b>Duration:</b>
<b>Complexity of case:</b> (circle one) <b>Standard</b> (TASC A or B) <b>Advanced</b> (TASC C or D)		

**The Trainee should explain what he/she intends to do throughout the procedure**  
**The Assessor should provide verbal prompts, if required, and intervene if patient safety is at risk.**  
**Rating: N = Not observed or not appropriate D = Development required      S = Satisfactory**  
**standard for CCT (no prompting or intervention required)**

	1.1.1      Competencies and Definitions	Rating N/D/S	Comments
	<b>Intra-procedural technique: global (G) and task-specific items (T)</b>		
IT1(G)	Follows an agreed, logical sequence or protocol for the procedure		
IT2(G)	Proceeds at appropriate pace with economy of movement		
IT3(G)	Anticipates and responds appropriately to variation e.g. anatomy		
IT4(G)	Deals calmly and effectively with unexpected events/complications		
IT5(G)	Uses assistant(s) to the best advantage at all times		
IT6(G)	Communicates clearly and consistently with nurses and radiographers		
IT7(G)	Ensures adequate radiation protection (including minimizing duration and dose of radiation by appropriate position and coning)		
IT9(T)	Inserts guidewire into CFA, introduces dilator and sheath, and removes dilator		
IT11(T)	Administers IV Heparin of appropriate dose		
IT12(T)	Advances guidewire across the lesion under fluoroscopic guidance		
IT13(T)	Check angio to ensure luminal position and confirm extent and position of disease		
IT14(T)	<b>Primary angioplasty</b> Introduces and advances angioplasty balloon of appropriate length and diameter to diseased section		
IT15(T)	Inflates and deflates balloon appropriately, and withdraws balloon after satisfactory inflation		
IT16(T)	Check angio to confirm adequate appearance and flow, and document run-off		

### Global summary

	<b>Level at which completed elements of the PBA were performed on this occasion</b>	<b>Tick as appropriate</b>
Level 0	Insufficient evidence observed to support a summary judgement	
Level 1	Unable to perform the procedure, or part observed, under supervision	
Level 2	Able to perform the procedure, or part observed, under supervision	
Level 3	Able to perform the procedure with minimum supervision (needed occasional help)	
Level 4	Competent to perform the procedure unsupervised (and could deal with any complications that arose)	

## Appendix 6- Modified Generic OSATS

Wire Handling					Score
1	2	3	4	5	
Awkward, tentative, inappropriate use, unaware of wire position, rarely maintains wire stability		occasionally stiff or awkward, or unaware of wire position, wire usually stable		fluid use of the wire, no awkwardness, aware of wire position, wire always stable	
Knowledge of wires and catheters					
1	2	3	4	5	
Frequently asked for the wrong wire or catheter or used inappropriate instrument		New the names of the wire or catheter and used appropriate instrument for task		Obviously familiar with wires or catheters and their names, and used the right instrument for the tasks	
Wire/ catheter technique					
1	2	3	4	5	
imprecise technique, frequent overshooting		precise technique; occasional overshooting		perfect precise technique	
Knowledge of specific procedure					
1	2	3	4	5	
Deficient knowledge, needed specific instruction at most procedural steps		Knew all important aspects of the procedure		Demonstrated familiarity with all aspect of the procedure	
Awareness of fluoroscopy and contrast use					
1	2	3	4	5	
Excessive use of fluoro and contrast		appropriate use, some unnecessary use		clear economy of fluoro and contrast; maximum efficiency	
Efficiency of Motion					
1	2	3	4	5	
Inefficient movements and many unnecessary moves		Efficient movements with some unnecessary moves		Highly efficient movement with very few unnecessary moves, clear economy of motion	
Flow of procedure					
1	2	3	4	5	
frequent pauses in procedure with uncertainty about next step		some pauses in procedure with little uncertainty about next step		well planned procedure with minimal uncertainty about next step	
Need for Verbal Prompts					
1		3		5	
Repeatedly needed prompts		Needed prompts sometimes		Able to complete the case without prompts	

## Appendix 7 Formative Supervised Surgical Assessment

Procedure: Sapheno-femoral junction ligation

1	2	3	4	5	Comments / Score out of 5
<b>Respect for Tissues</b>					
Handles tissue poorly causing frequent damage		Handles tissue carefully with occasional damage		Handles tissue well with minimal damage	<i>4- gentle use on tissue with minor exceptions.</i>
<b>Efficiency of Motion</b>					
Inefficient movements and many unnecessary moves		Efficient movements with some unnecessary moves		Highly efficient movement with very few unnecessary moves	<i>3- not yet as efficient as possible but will come with further practice</i>
<b>Instrument Handling</b>					
Repeatedly makes tentative or awkward moves with instrument		Competent use of instruments although occasionally appeared stiff or awkward		Fluid moves with instruments and no awkwardness	<i>4- holds and uses the instruments well</i>
<b>Use of assistants</b>					
Consistently placed assistants poorly or failed to use assistants		Good use of assistants most of the time		Strategically used assistant to the best advantage at all times	<i>2- the trainer did most of the assisting without input from the trainee</i>
<b>Knowledge of instruments</b>					
Frequently asked for the wrong instrument or used inappropriate instrument		Knew the name of most of the instruments and used appropriate instrument for task		Obviously familiar with the instruments required and their names	<i>4- clearly familiar with the instruments for this operation</i>
<b>Knowledge of specific procedure</b>					
Deficient knowledge, needed specific instruction at most operative steps		Knew all important aspects of the operation		Demonstrated familiarity with all aspect of the operation	<i>5-familiar with all the steps</i>
<b>Flow of operation</b>					
frequent pauses in operation with uncertainty about next step		some pauses in operation with little uncertainty about next step		well planned operation with minimal uncertainty about next step	<i>3- some pauses as unfamiliar as primary operator at the moment.</i>
<b>Total Score</b>					<b>22/35</b>

Level at which the completed supervised learning episode was performed on this occasion		Tick as appropriate
Level 1a	Able to assist with guidance (was not familiar with all steps of procedure)	
Level 1b	Able to assist without guidance (knew all the steps of procedure and anticipated next move)	
Level 2a	Guidance required for most/all of the procedure (or part performed)	
Level 2b	Guidance or intervention required for key steps only	
Level 3a	Procedure performed with minimal guidance or intervention (needed occasional help)	✓
Level 3b	Procedure performed without guidance or intervention but hesitant	
Level 4a	Procedure performed to a high standard without any guidance or intervention	
Level 4b	As 4a and was able to anticipate, avoid and /or deal with common problems or complications	

Degree of difficulty: easy/standard/difficult      Simulated: yes/no

Feedback after discussion with trainee:

**Operative Feedback**

***The trainee performed well on this occasion, and knew all the steps of the procedure. Needs to focus on directing their assistant to best advantage in this operation. Good instrument handling and knot tying laying the knots square and in the proper manner.***

***For next time try to be more assertive with your assistant in direction to best aid your exposure. More practice will increase your efficiency of motion.***

**Trainee Reflection** (e.g. The trainee should consider if their performance had improved since the last assessment for this operation and addressed any elements highlighted in previous assessments)

## Glossary of terms and definitions

**ADEPT**- Advance Dundee Endoscopic Psychomotor Trainer

**Anastomosis**- surgical term used to describe the artificial, often hand sewn join between two blood vessels, pieces of bowel, or organic and synthetic material.

**Angiogram**- an endovascular procedure where radio-opaque dye is injected into arteries or vein and an x-ray machine used to take a detailed image of the dye and hence the blood vessels

**ANGIOMentor**- A high fidelity electronic simulator which re-creates and replicates endovascular procedures authentically with haptic feedback

**Angioplasty**- an endovascular procedure to internally stretch arteries from the inside usually with a balloon

**ANOVA**- statistical test used to compare two or more groups of means in a parametric population of data.

**ARCP**- Annual review of competency progression: yearly structured interview process

**AoP**- Assessment of performance. A formative assessment of surgical skill which can inform the progression for a surgical trainee

**Basic surgical trainee**- surgical trainee with a minimum of 1 year post registration qualification and typically assigned a training grade of CT1/CT2 or ST1/ST2

**Bench Models**- small, synthetic, low fidelity models used in surgical skill labs which are designed to replicate part of an operations

**Box trainer**- often used in basic laparoscopic surgical skills to allow trainee to practice moving small items such as sugar cubes and needles and sutures around to develop the spacial awareness required before being allowed to practice such skills in the abdominal or thoracic cavities of real patients

**Cadaveric**- teaching material that is taken from donated cadavers, it may include part of, or all of a body.

**Carotid endarterectomy**- surgical procedure where the carotid artery in the neck is opened, hollowed out and a synthetic patch sewn over the opening to widen the artery and remove fatty deposits that can cause strokes.

**CAS**- Carotid artery stenting, a procedure to line the main carotid artery in the neck with a stent to hold a narrowed, diseased artery open and to prevent stroke.

**Catheter**- In endovascular surgery, a catheter is a long tube with a variously shaped end that can be used to deliver guide wires into blood vessels or inject dye

**CBD**- Case Based Discussion: Work placed based assessment which can be utilised to discuss and provide a record of and for learning of different clinical cases

**CCT**- certificate of completion of training: Qualification issued by the royal colleges after written and verbal examination after completion of speciality training

**CEX or Mini CEX- Clinical Evaluation Exercise**: Work placed based assessment which is used to discuss and provide a record of and for learning of a clinical encounter, typically history taking or examination skills

**Contrast**- the radio opaque dye used to inject or swallow in x-ray studies which is used to highlight anatomy for diagnostic purposes

**COPMED**- Conference of Postgraduate Medical Deans

**Coronary**- pertaining to the heart e.g. coronary arteries, the arteries surrounding the heart

**Cronbachs alpha**- a hypothesis free estimation of reliability, utilising classical theory, which generates reliability co-efficient between 0-1

**DOPS**- Direct observation of procedural skill: Work placed based assessment used to record and discuss progress in a minor procedural skill or part of an operation

**D-study**- A statistical method used to determine reliability, it aims to analyse and quantify all sources of variance which may affect reliability of an instrument within a study so that the result may be generalised. Like many forms of reliability it generates a reliability co-efficient between 0-1

**Endoscopic-** Term typically used to describe any procedure that involves a camera or video image used to investigate part of the body such as the stomach, bowel or bladder

**Endovascular-** used to describe vascular procedures that are achieved through minimally invasive access to the body via the arteries and the veins

**EWTR-** European working time regulation

**Fluoroscopy-** Continuous used x-ray which provides moving x-ray images for diagnostic and therapeutic purposes.

**Generalizable theory (GT)-** a statistical framework, which is hypothesis free, which allows estimation of reliability, where there are multiple sources of variance typically used to estimate reliability in performance assessments.

**GOALS-** Global observational assessment of laparoscopic skills, an assessment tool for laparoscopic procedures which requires a rater

**GSS- Global summary score of PBA:** the second part of the PBA form which is marked from 0-4, each item of which is defined by explicit descriptors

**GTSC- Global and task specific checklist of PBA,** which is part five of the six domains found in the first part of the PBA. Within this section each item is classified as either global or task specific to the procedure.

**Haptic feedback-** description applied to the tactile feedback given by high fidelity simulators, designed to mimic the feel of real tissue on the simulated instruments

**High fidelity-** a simulation scenario or simulator which aims to re-create a learning opportunity which is as close to real life as possible

**HUESAD-** Hiroshima University surgical assessment device- a motion tracking device

**ICEPS-** Imperial College Evaluation of Procedural Technical Skills, an index specific assessment tool for sapheno-femoral junction ligation

**Iliac artery-** one of the paired main arteries that supplies blood to the pelvis and unilateral leg

**Interventional radiologist-** Radiologist who utilises imaging media such as x-rays and ultrasound to perform invasive procedures on patients with diagnostic or therapeutic intent

**ISCP**- Intercollegiate surgical curriculum programme: the web based curriculum and surgical log book which all surgical trainees must use to store their log book experience and record evidence of work-placed based assessments

**KW- Kruskal Wallis** test- a non parametric test to determine the difference between multiple unrelated groups of different sizes

**Laparoscopic**- Term typically used to describe keyhole surgery in the abdomen or thoracic cavity which the operator relies on a video camera to transmit the images of the operation

**Logbook**- in surgical terms the number and variety of operations performed by a surgeon through their training career

**Low fidelity**- a basic simulation scenario or simulator which re-creates a learning opportunity that mimics real life but does not aim to reproduce real life

**MIST-VR**- Minimally invasive surgical trainer- Virtual Reality.

**MSF**- Multi source feedback: the process of triangulating behavioural reports from multiple people on a single individual's behaviour

**MWU**- Mann Whitney U- a non parametric statistical test which compares the means of two related samples

**New Deal**- Legislation which pre-dated the EWTR to restrict doctors working hours and ensure fairer pay for extra hours worked

**OCAP**- Orthopaedic Competency Assessment Project

**OCHRA**- Observational Clinical Human Reliability Assessment, an assessment method, utilising video analysis and raters to assess performance

**OSATS**- Objective Structured Assessment of Technical Skills: an assessment tool used to objectively assess surgical trainees in simulated and live surgical procedures

**OSCE**- objective structured clinical examination: simulated encounters or demonstration of clinical or surgical skills which are assessed by experienced assessors across a range of stations to assess medical students or trainees

**Patch angioplasty**- surgical procedure of widening an arterial by sewing a patch into a blood vessel to widen that area.

**PBA**- procedure based assessment: the main work placed based assessment used by surgical trainees to record progress in surgical psychomotor skills

**ProMIS**- Trade Name of a Laparoscopic Surgical simulator which contains validated training programmes

**Psychometrics**- the discipline of quantifying and measuring human behavioural performance

**Psychomotor**- term used to describe the integration of cognitive and motor skills

**Rater**- the term assigned to the individual assessor using the assessment form

**Sapheno-femoral junction ligation**- the surgical procedure where the superficial venous system is disconnected from the deep venous system in the groin, typically performed to treat varicose veins

**SF36**- Short form 36. A generic quality of life questionnaire

**SFA**- Superficial femoral artery, one of the paired arteries that supply the unilateral leg

**Spearman Rank Correlation Co-efficient**- Non parametric statistical test which compares correlations between two different but related data sets based on their rank order rather than statistical means

**SLE**- Supervised Learning Event. A formative assessment process which allows the trainee to reflect on their performance and provide feedback for a future event

**Surgical registrar**- grade of surgical trainee, typically assigned a training grade of ST3-ST8 in the UK.

**Suturing Jig**- typically a small, low fidelity platform which can anchor synthetic tissue or blood vessels in place which allows trainees to practice based surgical skills.

**Tibial artery**- one of the three small arteries that are found in the lower leg

**TrEndo-** Tracking Endoscopy- a motion tracking device

**Vein patch-** a synthetic or small piece of vein used to insert into an artery to prevent narrowing or deliberately widen an artery

**WBA-** Work placed based assessments: an electronic or hard copy record of a trainee interaction with a trainer to provide a record of and for learning

**WSR-** Wilcoxon signed Rank Sum test- to determine the mean difference in two or more related samples in a non parametric population.



Filename: MD Thesis final print run.docx  
Directory: C:\Users\Lucy\Documents  
Template: C:\Users\Lucy\AppData\Roaming\Microsoft\Templates\Normal.dot  
m  
Title:  
Subject:  
Author: Lucygreen  
Keywords:  
Comments:  
Creation Date: 29/07/2015 10:32:00  
Change Number: 6  
Last Saved On: 29/07/2015 11:28:00  
Last Saved By: Lucygreen  
Total Editing Time: 53 Minutes  
Last Printed On: 29/07/2015 11:28:00  
As of Last Complete Printing  
Number of Pages: 202  
Number of Words: 100,221 (approx.)  
Number of Characters: 571,261 (approx.)