The Bone & Joint Journal

Journal Club: 4 April 2014
Chairman: Prof Maxim Fehily
Attendees: Mr John Henderson, Mr Philip Wykes, Mr James Warner, Mr Alun Wall, Miss Victoria Sinclair, Mr Jon Baxter, Mr David Hawkes, Mr Jay Watson, Mr Mounir Hakimi, Mr Ricci Plastow, Dr Al-Asadi, Janice Taylor, Dr Alex Sewell, Ms Sue Greenhalgh, Ms Fiona Wardle, Ms Lynne Ronan, Ms Angela Shore, Mr Brian Donaldson, Mr Ian Joynes and Mr Andrew Maskell
Royal Bolton Hospital Orthopaedic Journal Club

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 Technol Assess* 2011;15 (1)

Reviewers: Jay Watson and Ricci Plastow

Why was the study done? What question are they addressing?
The Postgraduate Medical Education and Training Board requires postgraduate medical specialties to provide comprehensive curricula, with syllabus defined competencies blue printed to an assessment system. The European Working Time Directive limits training time and as a result, surgical training opportunities need to be maximised. Work Based Assessments (WBAs) are one such opportunity. They aid learning, provide a surgical trainee with feedback and also have a summative role.

The aim was to compare three WBA methods of assessing the surgical skills of trainees, taking into account their educational impact. The aim was ultimately attained by reviewing the user satisfaction, acceptability, reliability and validity of assessments, across a range of different surgical specialties and index procedures.

Valid model?
The study undertaken was prospective, observational and took place over a two year period, within the operating theatres of three teaching hospitals in Sheffield.

Assessors varied from consultant surgeons and anesthetists to scrub staff, surgical care practitioners and research staff. Assessments reviewed Obstetrics and Gynaecology (O&G), Orthopaedic, General Surgery and Cardiac surgery trainees in the operating theatre environment. Technical and non-technical skills were assessed for each trainee on at least two occasions with a minimum of two assessors present. Structured questionnaires were subsequently completed by both the assessors and those being assessed. The reliability of each method was estimated using Generalisability theory.

WBA Assessment Methods Used:

- Objective Structured Assessment of Technical Skills (OSATS) used for Obstetricians and Gynaecology trainees only
- Procedure Based Assessments (PBA) to predominantly assess technical skills
- Non-technical Skills for Surgeons (NOTSS) for situation awareness, communication and teamwork, decision-making and leadership.

Results

Recruitment:

437 patients were included in the study from a total of 832 that were approached as suitable training cases. Reasons for non-recruitment included a lack of inpatient beds, lack of trainee available to perform the case (25% of those consented), a supervisor performing the case despite a trainee presence (20% of those consented) and no list time for training (12% of those consented).

Assessors:

51 consultant supervisors, 56 anesthetists, 39 scrub nurses, 2 surgical care practitioners, 4 independent undertook 1635 assessments on 85 trainees. The assessments can be broken down as follows: 749 PBA, 695 NOTSS and 191 OSATS.

Reliability:

The reliability of each WBA, was estimated using Generalisability theory. This estimate is based around the theory that none of the variation across scores is random, but all the variance can be attributed to one or other factor (e.g. the stringency of the assessor, the ability of the trainee, etc.). To produce dependable reliable estimates, it is essential to sample each relevant factor, principally trainees, cases and assessors, as widely and representatively as possible. An overall target of 450 cases was set. Using this theory, a score is formed ranging from 0-1, the closer to 1 the more reliable. The results were as follows:

- PBA possessed high reliability ($G>0.8$ for three assessors judging one different case each). The reliability of PBA was less for O&G trainees than non-O&G trainees.
- OSATS was less reliable than PBA ($G>0.8$ for five assessors judging one different case each).
- NOTSS demonstrated the lowest reliability ($G>0.8$ for six assessors judging one different case each).

Construct Validity:

The extent to which each WBA and its individual components tested the professional constructs on which they are based, was assessed for design and validity purposes.

- PBA construct validity was demonstrated by the significant correlation of scores with age, specialty training level, total years training and total experience in index procedure.
- OSATS did not demonstrate any evidence for construct validity.
- NOTSS assessments demonstrated construct validity including decision making, situation awareness, communication and leadership training years and experience in recent experience in index procedure.

Variation:

The sampling strategy utilised was designed to allow the estimation of variation in trainee performance between individual cases and types of index procedure and differences in case complexity, as well as variability in assessor stringency and subjectivity.

In terms of scoring between different designations of assessor, PBA & NOTSS had little variation between scoring, however OSATS had wide variation.

User Satisfaction Questionnaires:

- PBA provided predominantly positive responses, particularly from trainees. They felt its best use was as a feedback tool, secondly as an aid in learning and thirdly for summative reasons.
- OSATS had positive responses from assessors but predominantly negative from O&G trainees. From both groups, the greatest number of negative responses regarded its summative use.
- The scrub team felt the NOTTS assessment allowed them to judge interpersonal skills, communication, teamwork and leadership at 92%; cognitive skills e.g. decision making at 85%. There was less agreement from anaesthetists. 27% compared to the latter's 60%.

Conclusion

The PBA tool possesses good overall utility as an assessment method given the good evidence for high reliability, validity and user satisfaction/acceptability. OSATS is a less reliable method than PBA although good reliability (G>0.8) remains achievable using feasible numbers of assessor judgments.

Whether PBA or OSATS are used to assess surgical skills within a training programme, the purpose, timing and frequency of WBA requires detailed guidance for both trainees and clinical supervisors to ensure that they are used correctly and provide maximum educational effectiveness.

Lower numbers of assessments may be appropriate for summative reasons, particularly for PBAs, however it must be remembered that the primary purpose for these assessments is learning, which

requires frequent assessment. Furthermore user satisfaction for summative reasons in this paper is lower and must be addressed.

NOTSS is promising but needed greater numbers for reliability. This WBA be completed by various assessors without intensive assessor training. The suggestion is that surgical specialities may wish to consider including NOTSS or elements of it into their assessment framework.

The paper concludes by recognising its own recruitment difficulties, obstacles it encountered to training and methods for addressing these. Primarily it suggests removing trainees from on-call night duty and reserving more theatre time per case to facilitate training. Finally it is suggested that If the identified obstacles to recruitment were addressed, trainees might gain access to at least twice as many training cases within the same timeframe. This is very significant given the requirement for surgical training to be more efficient within shorter training schemes with fewer hours for training

Journal Club opinion

This is an interesting study that is relevant to the Orthopaedic unit at the Royal Bolton Hospital and provides a positive contribution to the medical education literature. It raises the issue of surgical training changing from apprenticeship to curriculum and assessment based achievements at the Annual Review of Competency Progression. It also acknowledges shorter duration of training time now available and UK NHS service pressures.

The paper reviews modern and more established assessment methods, but does not neglect non-technical skills. It is vital to know if the assessment tools reviewed are valid, reliable and aid educational impact in the context of modern day training. Although this paper does not clearly prove this, it presents their value relative to each other in light of the demand for the development of more efficient surgical training methods, in which supervised training opportunities are maximised.

Strengths:

- A comprehensive prospective study involving a spread of surgical specialities
- Recruitment difficulties highlight barriers to training to be addressed
- The use of anaesthetic consultants and scrub staff for the assessing surgical training was a novel concept. Our journal club agreed that scrub staff could provide valuable feedback to surgical trainees and advocate such a use.
- Relative reliability of different assessments determined to help assess relative value
- The detailed parent document to this executive summary is freely available

Limitations:

- Recruitment issues were present. Large numbers achieved but nonetheless a very poor response rate from the total approached.
- No hypothesis tested and so no power calculation was possible. It was not clear where the paper's target number of assessments was derived from.

- Low numbers of independent assessors (four) and surgical care practitioners contributed data for analysis.
- O&G assessments proved of lower validity and had greater variation in scores. The reasons for this were not clearly assessed by this study. The variation was in part attributed to the higher proportion of O&G trainees with training concerns (42% vs. 4%).
- Reliability estimates were complicated and due to the lack of raw data, difficult for the reader to validate.
- Many recommendations that have been proposed to improve surgical training opportunities are difficult to adopt for practical reasons, implementation relies on sacrificing the primary aim of service delivery.

LeBlanc J, Hutchison C, Hu Y, Donnon T. Feasibility and fidelity of practicing surgical fixation on a virtual ulna bone. *J Can Chir* 2013;56:E91-97.

Reviewer: Mr David Ball

Why was the study done? What question are they addressing?

Previous research has shown surgical simulators are useful tools and provide a safe environment for trainees to learn and practice their psychomotor skills. Surgical simulators with proven high fidelity can be incorporated into surgical curricula. The gold standard of simulation in orthopaedic surgery is the use of sawbones. The 'fidelity' of a simulator is the extent to which its appearance and behaviour match that of the real environment, such that the trainee can transfer their learning to the real environment. The purposes of this study were firstly to assess whether a newly developed virtual simulator with haptic technology has comparable levels of fidelity to trainees performing ulna ORIF on sawbones, using a reliable fidelity questionnaire. The aim was to assess the simulator's feasibility, i.e. cost effectiveness, by performing a cost analysis.

Valid model?

Trainees selected for the study were from a single programme. Informed consent and ethical approval was gained. The surgical simulator was developed in house at the institute of the study (University of Calgary, Alberta, Canada). The fidelity questionnaire developed was derived from a literature review of medical questionnaires used for simulation and the questions were designed to assess multiple facets of fidelity including environmental, equipment & psychological fidelity. Stratified randomisation of trainees was employed, stratifying them by postgraduate year and sex, and then randomly assigning trainees to begin with either the simulator or sawbones. Trainees had 10 minutes to familiarise themselves with tools, equipment, control features, etc. for both modes of simulation training. Trainees completed the fidelity questionnaires immediately after completion of each task. Statistical analysis in the form of a reliability coefficient was used for each fidelity domain

and for overall fidelity. Effect size differences were calculated and categorised as either small, medium or large effect. The fidelities of the three domains were compared between both sawbones and the new simulator but also between junior and senior trainees.

Fidelity measured via fidelity scores derived from questionnaires completed after use of both the virtual simulator and sawbones. Feasibility assessed by cost comparison of both simulation models.

Results

Twenty-two out of twenty-six trainees in the programme tool part, 11 were randomly assigned to each group.

Significantly higher scores of overall fidelity were reported by junior than senior trainees, with a large effect size difference.

In all three domains of fidelity the mean scores were significantly higher for sawbones than the simulator. However, the simulator showed no significant difference between mean scores for each of the three fidelity domains, whereas there was a significant difference between the psychological fidelity for sawbones compared with the environment and equipment fidelity domains.

The cost comparison revealed the start up costs of the simulator to be $85000 versus $4225 for the sawbones.

The annual costs for running the simulator versus sawbones (i.e. replacement of one bone per year per trainee and maintenance/replacement of equipment) were essentially equivocal, at approximately $4000 per annum.

Conclusions

The level of fidelity is significantly higher for sawbones than for a virtual simulator, particularly for psychological fidelity. Junior or novice trainees found virtual simulation more useful than senior trainees. The goal of this study was not to try to replace existing simulation models but to use virtual simulators as an additional resource/ training tool. Once established the annual costs of running either simulation are essentially the same.

Strengths:

- It is the first study to assess the fidelity of sawbones & is of overall good quality
- It is the first study to develop a questionnaire specifically assessing surgical fidelity in simulation and so its use can be incorporated by future studies to determine the usefulness and validity of various surgical simulators
- It involved stratified randomisation to allocate its trainees to each group
- Statistical analysis showed significant differences for fidelity between simulators
- No bias or conflict of interest evident.

Limitations:

- Sawbones were mentioned as the gold standard of simulation but our attendees suggested that cadaveric simulation is probably the gold standard and likely has the highest fidelity of all simulation models in this context.
- Trainees had only 10 minutes to familiarise themselves with the new virtual simulator but many would already have had experience with sawbones so familiarity may have altered the trainees' perception of the level of fidelity.
- One institution with only a small number of trainees used
- It was a single 'snapshot' of fidelity of simulation. It did not assess how many trainees had used sawbones as part of simulation previously.
- Only one procedure could be performed, i.e. ORIF ulna.

Journal Club opinion

This study provides a valuable contribution to our evidence base for medical education in orthopaedic surgery. There has been no previous development or use of virtual simulators with haptics that allow practice of fracture fixation techniques. Our recommendations for improvement and future study would be a multicentre study with a greater number of trainees, prior and more prolonged exposure to the virtual simulator, multiple procedures to cater for all levels of trainees, and longer duration of follow up with regular use of both simulation tasks might yield further differences in fidelity, particularly the assessment of cumulative fidelity, and to assess the learning curve with multiple uses. The final proposal for further study was to compare the feasibility and fidelity of these models with cadaveric simulation.

Yehyawi TM, Thomas TP, Ohrt GT, Marsh JL, Karam MD, Brown TD, Anderson DD. A simulation trainer for complex articular fracture surgery. *J Bone Joint Surg[Am]* 2013;95-A:e92

Reviewers: Usman N Bhatty and Bilal Barkatali

Why was the study done? What question are they addressing?

The introduction of working time directives has significantly impacted medical professionals, particularly surgical trainees who rely heavily on apprentice-styled teaching. In order to redress this balance, there has been much research into alternative forms of teaching such as virtual simulators. The authors propose a lab-based simulation to supplement current orthopaedic training.

3 very clear objectives were specified; namely to create a model on which such simulated training can take place, to objectively assess trainees undergoing this simulation (similar to the more familiar 'OSCE' assessment) and to verify this form of training by comparing the results of junior trainees to

their senior colleagues. The hypothesis being that senior trainees score better than their junior counterparts.

Valid model?

The fracture chosen to repair was a distal tibia plafond fracture. A high density polyurethane foam tibia replica was chosen due to its similarity, and when fixed with barium sulphate bears similar resemblance, to human bone. The distal anatomy of the tibia replica was modelled from CT views of normal anatomy, using 'rapid machining technology'. Later a 7.5kg mass was dropped from 50cm in a drop tower on to the tibia replica using a talus surrogate, in order to produce fracture fragments. These were then placed in a displaced configuration by a surgeon to recreate distorted human anatomy. A PMMA impression of the altered articular surface was produced, and its fracture fragments negatively cast, in order to produce identical copies. Subsequent casted tibias were laser scanned and compared to the original tibia, confirming almost perfect copies. The tibia from a Sawbones foot and ankle model was substituted for the new fractured distal tibia, using a posterior approach.

5 senior (year 4 or 5 trainees) and 9 junior (year 1 or 2) willing residents were non-randomly selected from the local population of trainees. They were provided with X-Rays and 3D CT reconstructions of the fractured tibias 10 minutes before the simulation. 30 minutes were allocated to perform a reduction and fixation (using K-wires and fluoroscopy). Their attempts were recorded on video camera. The time taken to completion and the accuracy of their reduction were measured; the reductions were laser scanned and compared to the original giving a value for the average articular surface error. Additionally hand motions were tracked using 4 sensors and values provided for the total distance their hands travelled as well as the number of times their hands changed direction. The videos were reviewed to confirm the trainees had completed certain pre-defined steps. A surgeon also subjectively scored their attempts giving a score of 0, 3 or 5 across a series of parameters. These were preparation for procedure, respect for tissue, time and motion, instrument handling, use of fluoroscopy, use of k-wires, procedural flow and knowledge, and overall performance. Student t-test was used to analyze the results, with $p < 0.05$.

It should be noted that a conflict of interest was declared; namely a financial association between one or more of the authors and a biomedical entity which 'could be perceived to influence what was written' in the study.

Results

Measure	Senior Residents	Junior Residents	P Value
Articular error/mm	3.00 ± 0.43	3.09 ± 1.25	0.86
Time to complete/min	13.43 ± 4.68	14.75 ± 7.78	0.73
Cumulative hand distance/m	79 ± 48	390 ± 176	<0.01
Discrete hand motions	540 ± 303	511 ± 227	0.88
Global rating score	3.20	2.57	0.27

Note only the cumulative hand distance was statistically significant (senior residents moving less distance). No significant differences were noted with respect to the procedural checklist.

Journal Club opinion

We agree with the authors in that there is much scope for simulated training in the future of surgical training. As stated in the introduction, there are some simulators already in use, notably LapTrainers.

Fixation of tibial plafond fractures is a common scenario and a shared skill across all trainees. However, perhaps a distal radius fracture would have been more appropriate as k-wires often form part of the operation itself (rather than being used to fix the fracture for the purposes of later analysis). Although the description of how tibia models were constructed was thorough (and would seem to work), it seems rather long-winded. We wonder whether 3D printing has a large role to play here in the future. Additionally the use of ambiguous terms such as 'moderately' displaced fracture fragments, reduces the reproducibility of this study.

Inserting the fractured tibias into a Sawbones model is a novel way of creating a simulated scenario. However, as mentioned in the paper, this would still not truly mimic the soft tissue of distal tibias (the neurovascular and musculoskeletal structures would of course be absent). Likewise there is an argument that osteopenic bones are more likely to suffer fractures in the civilian population; the bone model used does not account for this. Within our training programme, 3D CT reconstructions are not provided pre-operatively for distal tibia fractures.

The results of the study are clearly limited for a number of reasons. The authors acknowledge the small and non-randomized sample is the chief reason why most data analysed was not statistically significant. Despite this, they interpreted the smaller cumulative hand distance of more senior trainees as being more purposeful movement. Furthermore the authors stated more purposeful movement is less likely to lead to soft tissue damage. This is a fair observation and would certainly seem logical. However this finding is at odds with the overall articulation error; smaller cumulative hand distance has not resulted in smaller articulation error. Indeed if the study was larger, it may be argued that more purposeful movement does not directly impact upon overall reduction of the fracture.

Though, as raw data, the senior trainees generally scored better than junior trainees. As mentioned in the paper, perhaps if the sample size was larger these differences may have become statistically significant.

The authors interestingly noted little correlation between global rating score and duration, articular reduction error and number of discrete hand motions. This would suggest a difference between what is subjectively seen as a good reduction and the objective markers provided.

The authors stated 3 aims. We feel a physical model was developed that could be used by centres elsewhere, but perhaps 3D printing may offer an easier alternative. Although an objective assessment of technical skill was performed (and may be reproduced), we felt there was more of a focus upon subjective markers (which did not correlated with overall articular reduction). Generally senior trainees performed better than their junior colleagues, but few statistically significant results were noted.

Overall this paper highlights a promising future for simulated training within orthopaedic surgery, and serves as a good pilot study for more thorough research.