BJO

■ FOOT & ANKLE

Assessing technical skill in ankle fracture surgery from the postoperative radiograph

PILOT DEVELOPMENT AND VALIDATION OF A FINAL PRODUCT ANALYSIS CORE OUTCOME SET

H. K. James,
J. Griffin,
G. T. R. Pattison

From Warwick Clinical Trials Unit, Coventry, UK

Aims

To identify a core outcome set of postoperative radiographic measurements to assess technical skill in ankle fracture open reduction internal fixation (ORIF), and to validate these against Van der Vleuten's criteria for effective assessment.

Methods

An e-Delphi exercise was undertaken at a major trauma centre (n = 39) to identify relevant parameters. Feasibility was tested by two authors. Reliability and validity was tested using postoperative radiographs of ankle fracture operations performed by trainees enrolled in an educational trial (IRCTN 20431944). To determine construct validity, trainees were divided into novice (performed < ten cases at baseline) and intermediate groups (performed ≥ ten cases at baseline). To assess concurrent validity, the procedure-based assessment (PBA) was considered the gold standard. The inter-rater and intrarater reliability was tested using a randomly selected subset of 25 cases.

Results

Overall, 235 ankle ORIFs were performed by 24 postgraduate year three to five trainees during ten months at nine NHS hospitals in England, UK. Overall, 42 PBAs were completed. The e-Delphi panel identified five 'final product analysis' parameters and defined acceptability thresholds: medial clear space (MCS); medial malleolar displacement (MMD); lateral malleolar displacement (LMD); tibiofibular clear space (TFCS) (all in mm); and talocrural angle (TCA) in degrees. Face validity, content validity, and feasibility were excellent. PBA global rating scale scores in this population showed excellent construct validity as continuous (p < 0.001) and categorical (p = 0.001) variables. Concurrent validity of all metrics was poor against PBA score. Intrarater reliability was substantial for all parameters (intraclass correlation coefficient (ICC) > 0.8), and inter-rater reliability was substantial for LMD, MMD, TCA, and moderate (ICC 0.61 to 0.80) for MCS and TFCS. Assessment was time efficient compared to PBA.

Conclusion

Assessment of technical skill in ankle fracture surgery using the first postoperative radiograph satisfies the tested Van der Vleuten's utility criteria for effective assessment. 'Final product analysis' assessment may be useful to assess skill transfer in the simulation-based research setting.

Cite this article: *Bone Jt Open* 2022;3-6:502–509.

Correspondence should be sent to Hannah K James; email: hsmith22@doctors.org.uk

Keywords: Surgical education, Training, Assessment, Outcome validation

Introduction

The use of 'final product analysis' (FPA) to objectively and quantitatively assess technical skill in surgery is an emerging research area in surgical education.[1] FPA is the technique of scoring performance based on the quality of the final surgical 'product',[2] the most obvious example of which, in orthopaedic surgery, is the postoperative radiograph. FPA is an attractive candidate outcome measure for simulation-based studies where assessment of skill transfer from the laboratory to real-world practice is required. It may also have applicability in the modern competence-based training climate,[3] where the procedure-based assessment (PBA) has recognized limitations.[4,5]

Assessment of technical skill in the real-world clinical environment has significant known methodological challenges, a recent large systemic review[1] showed that none of the technical skill assessment tools currently in use in orthopaedic training across the world satisfy the Norcini criteria[6] for effective assessment. FPA is promising as it has been shown in the laboratory to have face,[7] content,[2,7] construct,[2,8,9] and concurrent[7] validity and even educational impact[2,10] across a wide variety of procedures, sub-specialities and learner experience levels. There is real-world evidence for the utility of FPA in the assessment of technical performance in hip fracture surgery in actual patients (dynamic hip screw and hemiarthroplasty),[11] but not for ankle fracture open reduction internal fixation (ORIF) surgery, which is another common indicative[12] procedure performed by orthopaedic trainees.

Unlike for hip fractures, there is a striking paucity of literature evidence on the relationship between postoperative radiological appearance and patient outcome following ankle ORIF fracture surgery. What evidence exists relates to the precision of the reduction post-surgery, as measured by medial clear space (MCS),[13] tibiofibular clear space (TFSC),[14] medial malleolar displacement (MMD), lateral malleolar reduction (LMR) accuracy,[15] and talocrural angle (TCA).[13,16] The most recent of this evidence is nearly 30 years old, suggesting it is a neglected research area.

Postoperative radiographs are used in everyday clinical practice to judge the success of ankle ORIF surgery and inform decisions about patients' further management. In the absence of accepted criteria to define technical success, this judgement appears to be made using a global, qualitative 'expert eye' judgement, developed through experience.

The need to define a radiological FPA core outcome measurement set is driven principally by the requirement in the research setting for an outcome measure that is objectively measurable, valid, reproducible and reliable. Postoperative radiographs lend themselves conveniently to FPA as they are proximate to the time of surgery, collected as part of routine clinical care and neither invasive nor burdensome for the patient. The necessary properties of an FPA outcome set in ankle fracture ORIF surgery are, as for any similar procedure, that they are easily measurable on a standard anteroposterior postoperative radiograph, that they are clinically relevant, and responsive to changes in technical skill.

It is clear from searching the literature that there is no clear contemporary agreement on what constitutes a technically successful ankle ORIF. The primary objective of this study was to define a core outcome set of measurements obtainable from the postoperative radiograph to assess the technical success of an ankle fracture ORIF, and to assess their performance against four of Van der Vleuten's five criteria for effective assessment; validity, reliability, feasibility and cost-effectiveness.[17] We did not seek to test educational impact in this study.

Methods. Ethics permission for the main trial from which radiographs were analyzed for the purposes of this work was granted by the NHS Research Authority South Birmingham Research Ethics Committee (15/WM/0464) and the Confidentiality Advisory Group (16/CAG/0125). The e-Delphi exercise was run in tandem with another project assessing FPA for hip fracture technical performance and is reported in full elsewhere,[11] a summary of the methods are provided here. The study was conducted in three phases.

Phase one: consensus exercise to define core outcomes. A scoping literature search was conducted to identify evidence for FPA in assessing technical skill in ankle fracture ORIF. None was found, so the focus of the review was moved to look for evidence of radiological measurements as predictors of clinical outcome following ankle fracture surgery. A list of candidate measurement items was developed based on the scoping review and was externally checked with an internationally recognized expert in ankle fracture management. The Ankle Injury Rehabilitation trial radiology manual was consulted with permission as part of the scoping review.[18,19]

An e-Delphi exercise was undertaken to establish consensus on the FPA candidate items where no agreement exists. The consensus benchmark was set at ≥ 75% panel agreement, which is line with usual practice in the consensus setting literature.[20] The consultant orthopaedic surgeon body (n = 39) in a regional major trauma centre in the West Midlands, England, UK, were invited by email to participate. Overall, 19 participants (49%) completed the three survey rounds. A pragmatic decision to invite all consultants regardless of subspecialism was taken as ankle ORIF is an indicative surgical procedure that all trauma & orthopaedic consultants could reasonably be expected to manage as part of the unselected emergency take. The e-Delphi survey was hosted by an established online survey platform (Survey Monkey, USA). In round one, participants were presented with a list of candidate

Table I. Baseline characteristics of participating surgeons.

Variable	Novice (n = 17)	Intermediate (n = 7)
Mean age, yrs (range; SD)	28.7 (26 to 37; 2.9)	29.3 (26 to 35; 3.1)
Female, n (%)	3 (18)	3 (43)
Mean completed months of T&O training (SD)	11.4 (9.2)	22 (18.4)
Mean ankle fracture cases performed at baseline (range)	2.5 (0 to 8)	13.4 (10 to 17)

SD, standard deviation; T&O, trauma & orthopaedics.

Table II. Fracture complexity by surgeon experience in ankle procedures.

Fracture complexity*	Novice (residents < 10 cases)	Intermediate (residents ≥ 10 cases)	Total
Total, n	133	75	208
Simple, n (%)	113 (85)	67 (90)	180 (87)
Moderate, n (%)	6 (5)	3 (4)	9 (4)
Complex, n (%)	14 (11)	5 (7)	19 (9)

*See Supplementary Material for fracture classification system details.

FPA measurement items and given binary yes/no options as to whether they considered them appropriate for assessing technical skill in ankle ORIF. There was free text space for recording comments. In round two, items that had achieved consensus in round one were re-presented with proposed acceptability cut-off thresholds, again with binary yes/no agreement options. Some supporting literature evidence was provided as justification for the chosen levels. Free text space was again available for participants to expand on their answer. Items that had not reached consensus in round one (≥ 75% agreement) were re-presented in round two alongside additional supporting literature evidence summaries where appropriate. This process was repeated in round three. Items that had not reached consensus after three rounds were abandoned.

Phase two: feasibility testing. Feasibility of obtaining radiological measurements identified during phase one was assessed by two of the authors (GTRP, HKJ). Measurements were made using the inbuilt tools within the PACS (Picture Archiving and Communication System) user interface, which is the common clinical radiography software package used throughout the NHS. The first formal postoperative anteroposterior radiograph was used. Intraoperative image intensifier radiographs were not used for this study.

Phase three: validity and reliability testing. Radiographs of operations performed by trainee participants in a multicentre educational randomized trial (ISRCTN 20431944) were used for validity and reliability testing. The study protocol details participant demographics and eligibility criteria.[21] Face and content validity of candidate items was determined in phase one. There were no a priori inclusion criteria for ankle fracture severity or configuration. Radiographs were retrieved for all procedures that were coded in participating trainee logbooks as 'ORIF ankle, trauma'. We assigned all fractures an AO classification code based on the preoperative radiograph. These were then subsequently categorized into 'simple', 'moderate' and 'complex' fractures for the purposes of analysis to ensure that the case mix was balanced between groups

to ensure a fair comparison (see Supplementary Material for detailed classification matrix).

Construct validity, defined as the ability of an instrument to discriminate between skill levels, was measured by assessing novice and intermediate level trainee performance. For the purposes of this validation study, 'novices' were arbitrarily defined as having performed < ten ankle ORIF procedures at baseline (recruitment into the trial) and 'intermediates' were defined as having performed ≥ ten ankle ORIFs at baseline. There were no studies in the literature exploring the learning curve for ankle ORIF surgery, so a pragmatic decision was made to determine the cut-off at 10 cases.

Statistical analysis. For categorical outcomes, chi-squared tests of association were undertaken, and Fisher's exact test was used if the cell count was < five. For continuous outcomes, the means between groups were compared using independent-samples t-test, and tested whether the difference between groups was zero. Concurrent validity, defined here as being the performance of an assessment tool against the current 'gold standard', was examined by comparing the candidate radiological items against the PBA global scale score, which was considered in both categorical and continuous forms.

All primary measurements for the validity testing were taken by one author (HKJ). The advising statistician determined that 25 cases was an adequate reliability testing sample, hence a randomly selected subset of 25 cases was used. To determine intrarater reliability, these 25 cases were measured by the same assessor (HKJ) one week apart, and on one occasion by a second rater (GTRP) to determine inter-rater reliability. For categorical outcomes, Cohen's kappa statistic was used to assess agreement and the crude percentage agreement in absolute terms was presented.

Results
In total, 235 ankle ORIF procedures were performed by 24 postgraduate year three to five orthopaedic residents during the 2014 to 2015 training year, at nine NHS hospitals in the West Midlands, UK. There were 42 PBAs were completed for these procedures in the study population. Baseline demographics of the study participants are shown in Table I. Only procedures coded in the electronic surgical logbook as 'supervised – trainer scrubbed' and
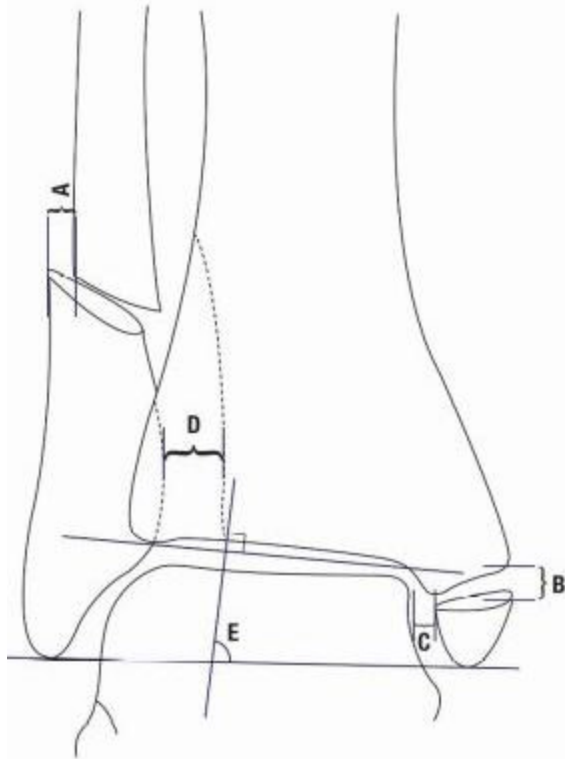
Fig. 1

Schematic diagram of final product analysis measurements. A) Lteral malleolus displacement (mm); B) medial malleolus displacement (mm); C) medial clear space (mm); D) tibiofibular clear space (mm); and E) talocrural angle (degrees).

Table III. Candidate item inclusion and acceptability threshold determination by Delphi round.

Item	Round in which consensus was achieved	Panel agreement, %
1a. Medial clear space	1	88
1b. Acceptable MCS ≤ 4 mm	2	76
2a. Medial malleolar displacement	1	96
2b. Acceptable MMD ≤ 2 mm	3	95
3a. Lateral malleolar displacement	1	96
3b. Acceptable LMD ≤ 2 mm	3	95
4a. Tibiofibular clear space	1	75
4b. Acceptable TFCS ≤ 5 mm	3	100
5a. Talocrural angle	3	95
5b. Acceptable = 80° ± 5	3	95
6a. Talar tilt angle	Failed to reach consensus after 3 rounds > excluded	
6b. Acceptable = N/A		

LMD, lateral malleolar displacement; MCS, medial clear space; MMD, medial malleolar displacement; N/A, not applicable; TFCS, tibiofibular clear space.

'supervised – trainer unscrubbed' were included. Ankle fracture complexity by group is shown in Table II.

Face and content validity. The face validity (that a tool appears superficially fit for purpose) and content validity (that an assessment tool tests appropriate/relevant domains) were demonstrated by externally checking the appropriateness of the initial list of candidate measurement items with an internationally renowned expert in ankle fractures, and by consultation with the AIR trial team. Five candidate FPA parameters achieved panel consensus > 75%: TCA (degrees), MCS (mm), MMD (mm), LMD (mm), and TFCS (mm). These are shown in schematic form in Figure 1. One measurement, talar tilt angle, was rejected as it failed to reach the 75% agreement threshold after three rounds, and was perceived as inferior to talocrural angle. The stage of achieving consensus, percentage agreement and acceptability thresholds for each candidate parameter is shown in Table III.

Construct validity. Construct validity, widely considered as the ability of an assessment instrument to discriminate between different levels of performance, was evaluated in this study by comparing between-group differences for the five FPA items. The groups were defined as novices

(performed < ten cases) or intermediate (performed ≥ ten cases) of ankle ORIF at ST-S/ST-U level at baseline.

All five items showed that performance was better in the intermediate group as compared to the novice group, and all except talocrural angle were statistically significant, suggesting excellent construct validity (Table IV). For MCS, the mean distance was 3.8 mm for novices versus 3.0 mm for intermediates ($p < 0.001$); for MMD, the mean distance was 1.8 mm for novices versus 0.8 mm for intermediates ($p = 0.001$); for LMD, the mean distance was 2.2 mm for novices versus 1.3 mm for intermediates (< 0.001); and for TFCS, the mean distance was 5.6 mm versus 4.1 mm ($p < 0.001$). The smaller distances seen on the radiographs of the cases performed by the intermediates as compared to the novices is indicative of a more precise reduction and hence technically superior performance.

The fracture complexity was assessed as simple, moderate or complex and compared between groups. It was found that the case-mix was well balanced between groups (Table II).

Concurrent validity. Concurrent validity of the FPA items was assessed by comparing the difference between the global PBA descriptor scores (contributed from across the study cohort) with each of the five radiological measurements. The PBA scores were considered as categorical variables (Table V). No significant association was seen between PBA score and any of the five FPA items using the chi-squared test.

Reliability. Intra-rater reliability was excellent for all five items (Table VI) (Cohen's kappa = 0.83, 0.93, 0.86, 0.85, and 0.81 for LMD, MMD, MCS, TFCS and TCA, respectively). Inter-rater reliability was found to be excellent

Table IV. Construct validity of ankle radiological outcome measures.

Outcome	Novice (residents < 10 cases)	Intermediate (residents ≥ 10 cases)	Total	p-value*
Total, n	147	88	235	
Talocrural angle, °				0.517
Number	124	83		
Mean (SD; range)	78.1 (3.5; 68.6 to 88.0)	77.8 (0.4; 70.3 to 87.7)	78.0 (3.4; 68.6 to 88.0)	
Medial clear space, mm				< 0.001
Number	130	83		
Mean (SD; range)	3.8 (1.3; 1.5 to 8.3)	3.0 (1.2; 0.1 to 8.6)	3.5 (1.4; 0.1 to 8.6)	
Medial malleolar displacement, mm				0.001
Number	130	83	213	
Mean (SD; range)	1.8 (2.3; 0 to 13.1)	0.8 (1.3; 0 to 7.3)	1.4 (2.0; 0 to 13.1)	
Lateral malleolar displacement, mm				< 0.001
Number	130	83	213	
Mean (SD; range)	2.2 (1.9; 0 to 7.8)	1.3 (1.4; 0 to 8.1)	1.9 (1.8; 0 to 8.1)	
Tibiofibular clear space, mm				< 0.001
Number	130	83	213	
Mean (SD; range)	5.6 (1.8; 2.0 to 10.7)	4.1 (1.4; 0.5 to 8.3)	5.0 (1.8; 0.5 to 10.7)	
PBA global rating scale (continuous)				< 0.001
Number	23	19	42	
Mean (SD; range)	2.1 (0.3; 2 to 3)	3 (0.5; 2 to 4)	2.5 (0.6; 2 to 4)	
PBA global rating scale (categorical), n (%)				0.001
2	21 (91)	2 (11)	23 (55)	
3	2 (9)	15 (79)	17 (41)	
4	0 (0)	42 (70)	2 (5)	

*Independent-samples *t*-test comparing means; chi-squared test where cells > five cases and, Fisher's exact test where cells ≤ five cases.
PBA, procedure-based assessment; SD, standard deviation.

Table V. Concurrent validity of ankle radiological outcome measures against procedure-based assessment (PBA).

Outcome	PBA score 2	PBA score 3	p-value*
Talocrural angle, °			
Number	21	17	0.894
Mean (SD; range)	77.4 (3.8; 70.3 to 84.0)	77.5 (3.2; 74 to 87)	
Medial clear space, mm			
Number	21	17	0.065
Mean (SD; range)	3.3 (1.1; 1.6 to 5.7)	2.7 (0.6; 1.7 to 3.8)	
Medial malleolar displacement, mm			0.677
Number	21	17	
Mean (SD; range)	0.9 (1.4; 0 to 4.1)	1.0 (1.1; 0 to 3.1)	
Lateral malleolar displacement, mm			0.482
Number	21	17	
Mean (SD; range)	1.8 (1.7; 0 to 4.9)	1.4 (1.5; 0 to 5.6)	
Tibiofibular clear space, mm			0.337
Number	21	17	
Mean (SD; range)	4.8 (1.9; 2.0 to 10.3)	4.3 (1.4; 1.8 to 7.0)	

*Independent samples *t*-test comparing means; chi-squared test where cells > fivecases; and Fisher's exact test where cells ≤ five cases.
SD, standard deviation.

for LMD, MMD and TCA (Cohen's kappa = 0.80, 0.95, and 0.80, respectively), and moderate for MCS and TFCS (Cohen's kappa = 0.52 and 0.60, respectively).

Cost effectiveness. Radiographs were taken as part of routine patient clinical care and hence did not incur an additional cost burden. It took the assessors a mean time of < two minutes to obtain the five measurements per patient. FPA may therefore be a cost-effective assessment method.

Discussion

The endpoint of surgical training is to produce competent surgeons who will perform safe, high-quality operations

Table VI. Intra- and inter-rater reliability of ankle outcome measures.

Outcome	N	Intra-rater (rater 1 vs rater 2), ICC (95% CI)	Inter-rater (rater 1 vs rater 2), ICC (95% CI)
Lateral malleolar displacement, mm	25	0.83 (0.66 to 0.92)	0.80 (0.60 to 0.91)
Medial malleolar displacement, mm	25	0.93 (0.84 to 0.97)	0.95 (0.89 to 0.98)
Medial clear space, mm	25	0.86 (0.70 to 0.93)	0.52 (0.16 to 0.75)
Tibiofibular clear space, mm	23	0.85 (0.68 to 0.93)	0.60 (0.25 to 0.81)
Talocrural angle, °	25	0.81 (0.62 to 0.91)	0.80 (0.59 to 0.91)

CI, confidence interval; ICC, intraclass correlation coefficient.

for their patients. It is widely recognized within the surgical community that assessing technical skill is methodologically[1] challenging.

Simulation based training in orthopaedics is becoming increasingly prevalent,[22] and there is a clear need for high quality evidence of skill transfer to the workplace. There is a widely held perception that educational assessment is somehow remote from, and independent of, clinical outcomes.[23] The ultimate goal of any educational intervention is to show benefit to patients, and hence the appeal of FPA is to marry the two; in being both robust as an assessment tool but also clinically relevant.

In this study, we have shown that FPA for assessing technical skill in ankle ORIF surgery is feasible, has face, content and construct validity, and is reliable. This satisfies the four domains of Van der Vleuten's utility criteria that we sought to test in this work. The educational impact potential of FPA requires separate investigation with different methodology.

Clearly, assessment of competency is far more complex and nuanced than simply examining the end-product of an operation, and we are not suggesting that FPA replaces the well-established and respected arsenal of educational assessment tools currently in use, such as the PBA and the new multi-consultant report. The role of FPA is likely to be most useful as an adjunct to existing assessment methods, with a particular value in the educational research setting where the impact of an educational intervention requires evaluation using objective, quantitative and reproducible outcome measures. It is highly likely that these individual five measurements are not useful in isolation, but would combination into a cumulative summary score, the domain weightings for which require determination in further work and are beyond the scope of this paper.

The strengths of this study are that we have systematically assessed 235 real operations performed by 24 trainees in nine regional secondary and tertiary hospital sites across one surgical training year, which is a much larger sample than is typically seen in educational studies. This suggests the generalizability of the results is good and the chance of a type-2 error minimized. We have followed a systematic framework for assessing the utility of our proposed FPA outcome set, using the Van der Vleuten's criteria for effective assessment, which is used across the medical education literature.

Weaknesses of this study are that, as discussed above, we have not explored the educational impact of the FPA outcome measures, which is arguably the most important dimension of any assessment tool. This will require further work, likely qualitative, to understand if/how feedback FPA influences technical performance.

As we evaluated radiographs from operations performed within the setting of an educational trial, we could not specify a priori which ankle fracture types were to be included in the analysis and this is a heterogenous injury. We made the pragmatic decision that inclusion of a breadth of trainee participant locations and experience levels (and by inference, the generalizability of the results) took priority over tight control of fracture type or surgical implant choice. For this reason, we did not include any implant-related FPA measurements in this study. Furthermore, in order to have utility as a patient-centred technical skills assessment tool, further work is required to examine the clinical relevance of the acceptability thresholds in terms of patient outcomes.

Similarly, by including nine NHS hospitals, there was likely to be some local practise variation in the method of obtaining the postoperative ankle radiographs (i.e. standard AP vs mortise, weightbearing vs non-weightbearing). Again, we felt the pragmatic approach of using images obtained in the natural clinical rather than research setting, with their inherently greater variability, improved the generalizability of our findings and applicability of our results to real-world educational research, potentially at the expense of an unknown degree of accuracy. The reliability of the measures was found to be high using ICC techniques, but it is important to acknowledge that this merely relates to the inter-rater and intrarater reliability of taking a set of measurements from digital radiographs. Any inferences about technical skill of the surgeon or outcome cannot be made from these results alone, as there are several other variables that contribute to assessment reliability, which were not tested here.

It was an interesting finding that there were between-group differences, despite all cases being supervised (coded in trainee logbooks as either supervised trainer-unscrubbed or supervised trainer-scrubbed), and this

finding is consistent with similar work by the authors examining FPA for hip fractures.[11] We do not have a ready explanation for this, and by not knowing with certainty what the clinical implications of these measurements are, we cannot speculate on whether they might impact the patient. Nonetheless, if we extend this further to say that we are seeing evidence of potentially technically inferior results in less experienced surgeons, then there are important issues to be discussed around operative autonomy at a junior training stage.

We hope both these works will prove useful to researchers who require a patient-centred outcome measure for measuring technical skill in orthopaedic surgery. They may also be useful in the clinical trial setting, when assessment of the quality of surgical fixation of ankle fractures is required. Of interest, concurrent validity with PBA was found to be poor across both studies. This is likely because they are measuring different aspects of competence and further reinforces the argument that FPA can augment rather than replace the higher order, holistic workplace assessment methods. Future research areas include development of a summative FPA score and assessment with a larger study cohort to see if surgeon experience relates to the expert-defined benchmarks in our Dephi study.

In conclusion, assessment of technical skill in ankle fracture surgery using the first postoperative radiograph satisfies Van der Vleuten's utility criteria for effective assessment; it is feasible, valid, reliable, and cost-effective. FPA assessment may be useful to complement the PBA assessment and has particular appeal for use in quantitatively assessing skill transfer in the educational and simulation-based research setting. Further work is required to explore the clinical relevance and educational impact of the FPA metrics.

Take home message

- Postoperative radiographs can be used to assess technical skill in ankle fracture surgery.
- This could be a useful adjunct to workplace based assessments, and has particular application to measuring skill transfer in the simulation research setting.

Twitter

Follow H. K. James @hannah_ortho
Follow Warwick Clinical Trials Unit @WarwickCTU

Supplementary material

Table showing the fracture and dislocation classification.

References

1. **James H**, **Chapman A**, **Pattison G**, et al. Assessment of technical skill acquisition and operative competence in trauma and orthopaedic surgical training: a systematic review. *JBJS(Am) Reviews*. 2019.
2. **Christian MW**, **Griffith C**, **Schoonover C**, et al. Construct validation of a novel hip fracture fixation surgical simulator. *J Am Acad Orthop Surg*. 2018;26(19):689–697.
3. **James HK**, **Gregory RJH**. The dawn of a new competency-based training era. *Bone Jt Open*. 2021;2(3):181–190.
4. **Davies RM**, **Hadfield-Law L**, **Turner PG**. Development and Evaluation of a New Formative Assessment of Surgical Performance. *J Surg Educ*. 2018;75(5):1309–1316.
5. **Hunter AR**, **Baird EJ**, **Reed MR**. Procedure-based assessments in trauma and orthopaedic training--The trainees' perspective. *Med Teach*. 2015;37(5):444–449.
6. **Norcini J**, **Anderson B**, **Bollela V**, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach*. 2011;33(3):206–214.
7. **Shi J**, **Hou Y**, **Lin Y**, **Chen H**, **Yuan W**, et al. Role of visuohaptic surgical training simulator in resident education of orthopedic surgery. *World Neurosurg*. 2018;111:e98–e104.
8. **Nousiainen MT**, **Omoto DM**, **Zingg PO**, **Weil YA**, **Mardam-Bey SW**, **Eward WC**, et al. Training femoral neck screw insertion skills to surgical trainees: computer-assisted surgery versus conventional fluoroscopic technique. *J Orthop Trauma*. 2013;27(2):87–92.
9. **Akhtar K**, **Sugand K**, **Sperrin M**, **Cobb J**, **Standfield N**, **Gupte C**, et al. Training safer orthopedic surgeons. Construct validation of a virtual-reality simulator for hip fracture surgery. *Acta Orthop*. 2015;86(5):616–621.
10. **Bergeson RK**, **Schwend RM**, **DeLucia T**, **Silva SR**, **Smith JE**, **Avilucea FR**, et al. How accurately do novice surgeons place thoracic pedicle screws with the free hand technique? *Spine (Phila Pa 1976)*. 2008;33(15):E501-7.
11. **James HK**, **Pattison GTR**, **Griffin J**, **Fisher JD**, **Griffin DR**, et al. Assessment of technical skill in hip fracture surgery using the postoperative radiograph: pilot development and validation of a final product analysis core outcome set. *Bone Jt Open*. 2020;1(9):594–604.
12. **No authors listed**. Joint Committee on Surgical Training. https://www.jcst.org (date last accessed 18 March 2022).
13. **Mont MA**, **Sedlin ED**, **Weiner LS**, **Miller AR**, et al. Postoperative radiographs as predictors of clinical outcome in unstable ankle fractures. *J Orthop Trauma*. 1992;6(3):352–357.
14. **Pettrone FA**, **Gail M**, **Pee D**, **Fitzpatrick T**, **Van Herpe LB**, et al. Quantitative criteria for prediction of the results after displaced fracture of the ankle. *J Bone Joint Surg Am*. 1983;65-A(5):667–677.
15. **Joy G**, **Patzakis MJ**, **Harvey JPJ**. Precise evaluation of the reduction of severe ankle fractures: technique and correlation with end results. *JBJS*. 1974;56(5):979–993.
16. **Phillips WA**, **Schwartz HS**, **Keller CS**, et al. A prospective, randomized study of the management of severe ankle fractures. *J Bone Joint Surg Am*. 1985;67-A(1):67–78.
17. **Van Der Vleuten CPM**. The assessment of professional competence: Developments, research and practical implications. *Adv Health Sci Educ Theory Pract*. 1996;1(1):41–67.
18. **Kearney RS**, **McKeown R**, **Stevens S**, et al. Cast versus functional brace in the rehabilitation of patients treated for an ankle fracture: protocol for the UK study of ankle injury rehabilitation (AIR) multicentre randomised trial. *BMJ Open*. 2018;8(12):e027242.
19. **Kearney R**, **McKeown R**, **Parsons H**, et al. Use of cast immobilisation versus removable brace in adults with an ankle fracture: multicentre randomised controlled trial. *BMJ*. 2021;374:1506.
20. **Diamond IR**, **Grant RC**, **Feldman BM**, et al. Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *J Clin Epidemiol*. 2014;67(4):401–409.
21. **James H**, **Pattison GTR**, **Fisher JD**, **Griffin DR**, et al. Cadaveric simulation vs standard training for postgraduate trauma & orthopaedic surgical trainees: protocol for the CAD:TRAUMA study multi-centre randomised controlled educational trial. 2020.
22. **James HK**, **Gregory RJH**, **Tennent D**, **Pattison GTR**, **Fisher JD**, **Griffin DR**. Current provision of simulation in the UK and Republic of Ireland trauma and orthopaedic specialist training: a national survey. *Bone Jt Open*. 2020;1(5):103–114.
23. **James H**. Measuring the educational impact of simulation training in trauma & orthopaedics. *Journal of Trauma & Orthopaedics*. 2019. https://www.boa.ac.uk/resources/knowledge-hub/jto-september-2019.html. In press.

Author information:
- H. K. James, PhD, MMedEd, MRCS, Clinical Science Research Laboratories, Warwick Clinical Trials Unit, Warwick Medical School, Coventry, UK; Department of Trauma & Orthopaedics, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK.
- J. Griffin, BSc, MSc, Research Fellow, Warwick Clinical Trials Unit, Warwick Medical School, Coventry, UK.

- G. T. R. Pattison, MSc (Med Ed), FRCSEd (Tr&Orth), Consultant Orthopaedic Surgeon, Department of Trauma & Orthopaedics, University Hospitals Coventry and Warwickshire NHS Trust, Coventry, UK.