Watson, M. J.; Wong, D. M.; Kluger, R.; Chuan, A.; Herrick, M. D.; Ng, I.; Castanelli, D. J.; Lin, L.; Lansdown, A. K.; Barrington, M. J. "Psychometric evaluation of a direct observation of procedural skills assessment tool for ultrasound-guided regional anaesthesia", Anaesthesia Vol. 69, Issue 6, p. 604-612 (2014)

Available from: http://dx.doi.org/10.1111/anae.12625

Assoc. Prof. Michael Barrington

Department of Anaesthesia
St. Vincent's Hospital, Melbourne
PO Box 2900
Fitzroy
Victoria 3065
Australia

Psychometric evaluation of a direct observation of procedural skills assessment tool for ultrasound-guided regional anaesthesia

Watson M.J.,[1] Wong D.M.,[2] Kluger R.,[3] Chuan A.,[4] Herrick M.D.,[5] Ng I.,[6] Castanelli D.J.,[7] Lin L.,[8] Lansdown A.K.,[9] and Barrington M.J.[10]

1 Medical Student, University of Newcastle, Newcastle, Australia.
2 Consultant Anaesthetist, St. Vincent's Hospital, Melbourne, Australia.
3 Senior Staff Specialist Anaesthetist and Supervisor of Training, St. Vincent's Hospital, Melbourne
4 Consultant Anaesthetist, Liverpool Hospital; Senior Clinical Lecturer, University of Melbourne; Senior Clinical Lecturer, Macquarie University
Macquarie University, Sydney, Australia.
5 Assistant Professor of Anesthesiology, Director of Regional Anesthesiology, Geisel School of Medicine at Dartmouth, Lebanon, USA.
6 Staff Specialist Anaesthetist, Royal Melbourne Hospital; Honorary Fellow, University of Melbourne, Australia.
7 Staff Specialist Anaesthetist, Monash Medical Centre, Department of Anaesthesia and Perioperative Medicine, Monash University, Melbourne, Australia
8 Consultant Anaesthetist, Austin Health, Melbourne, Australia
9 Consultant Anaesthetist, Royal Prince Alfred Hospital; Associate Clinical Lecturer, Sydney Medical School, Sydney Australia
10 Senior Staff Specialist Anaesthetist, St. Vincent's Hospital, Melbourne;
Associate Professor, Melbourne Medical School, Faculty of Medicine, Dentistry and Health Sciences, University of Melbourne, Australia.

Correspondence to: Assoc. Prof Michael Barrington (michael.barrington@svhm.org.au)

Psychometric evaluation of direct observation of procedural skills

Summary

Assessment tools must be investigated for reliability, validity and feasibility before being implemented. In 2013, the Australian and New Zealand College of Anaesthetists introduced workplace-based assessments, including a direct observation of procedural skills assessment tool. The objective of this study was to evaluate the psychometric properties of this assessment tool for ultrasound-guided regional anaesthesia. Six experts assessed 30 videoed trainee performances of ultrasound-guided regional anaesthesia. Inter-rater reliability assessed using absolute agreement intraclass correlation coefficients, varied from 0.10 – 0.49 for the 9 individual 9-point scale items and was 0.25 for a 'total score' of all items. Internal consistency was indicated by correlation between 'total score' and 'overall performance' scale item ($r = 0.68$, $p < 0.001$). Construct validity was demonstrated by the 'total score' correlating with trainee experience ($r = 0.51$, $p = 0.004$). The mean time to complete assessments was 6 minutes 35 seconds.

Introduction

The medical profession has long determined and regulated its own standards of excellence.[1] However, recent decades have seen increasing societal pressure for medical educators to validate their training methods.[1-4] Increased concern for patient safety has driven medical education toward ‘competency-based’ assessments, leading to a demand for reliable, valid and feasible methods of clinical skills assessment.[5-9] Reliability refers to the consistency and reproducibility of results produced by an assessment tool. External reliability (often measured by inter-rater reliability) measures the ability of a tool to produce consistent scores across a range of assessors. Construct validity refers to the ability of an assessment tool to differentiate between individuals with various levels of expertise.[5]

The Direct Observation of Procedural Skills (DOPS) assessment tool was introduced by the Royal College of Physicians in 2003 as one means of Workplace Based Assessment (WBA).[10] WBA involves ongoing formative assessment of knowledge, clinical decision making and procedural skills.[10] Between 2008 and 2012, the Australian and New Zealand College of Anaesthetists (ANZCA) undertook a review of their curriculum.[11] Commencing in 2013, DOPS assessment became an integral part of this revised training program[11, 12] DOPS has been modified by many medical colleges including ANZCA. The ANZCA DOPS uses behavioural descriptive anchors along a 9-point rating scale reflecting level of competence and need for supervision (Figure S1). To date, no study has specifically evaluated the psychometric properties of ANZCA DOPS for clinical anaesthetic skills.

Assessment tools should be rigorously investigated for reliability, validity, feasibility and comprehensiveness before being adopted for use. [5, 13, 14] The objective of this study was to evaluate the psychometric properties (inter-rater reliability, internal consistency, construct validity and feasibility) of ANZCA DOPS for ultrasound-guided regional anaesthesia (UGRA) skills.

Methods

The Human Research Ethics Committee of St Vincent's Hospital, Melbourne granted approval for this project (LRR protocol number 134/12) and to store data in a remote secure database through an online interface at www.anaesthesiaregistry.org (Protocol number: Q/A 039/08). Written informed consent for the video recording and assessment of procedures was obtained from all study participants (patients and trainees).

Trainees were filmed performing ultrasound-guided peripheral nerve blocks while being supervised by two authors (DW and MB) not involved with scoring. Inclusion criteria were patients receiving single injection ultrasound-guided peripheral nerve blocks placed using an in-plane needle technique. Block performance occurred in a context close to 'real-life practice'. Prior to filming, anaesthetic and nursing staff were informed of the purpose of the film and asked not to comment, intervene or prepare equipment without being prompted by the trainee. The supervising anaesthetist took ultimate responsibility for the procedure and ensured the patient's safety during block performance and intervened when necessary. Supervisors ensured patients were appropriately monitored, positioned and sedated throughout, providing standardised prompts when trainee's faulted during set-up, scanning and nerve block performance (Appendix).

A HDR-CX190 Handycam® (Sony Corp, Tokyo, Japan) and tripod were used to capture live video of the procedure and immediate surroundings. This external video was combined with concurrent recording of ultrasound images. A Sonosite© M-Turbo (Bothell, USA) ultrasound machine was used for all procedures with intermediate and high frequency linear probes. Ultrasound data was exported to iMovie using a Canopus©ADVC-55 digital video converter (Kobe-city Hyogo, Japan). Videos were edited on iMovie (Version 8.0.6, Apple Inc. Cupertino, USA) with ultrasound images displayed picture-in-picture and synchronised with external video. Videos included all relevant aspects of the procedure, room set-up and patient

and team interactions. Editing occurred to preserve patient's privacy and to remove footage not relevant to assessment such as intravenous cannulation.

Six assessors actively involved with training and supervision of registrars (authors: AC, LL, IN, MH, AL and DC) were trained in the use of DOPS during three iterative rounds of pilot use. Results of pilot assessment were tabulated, disseminated and scrutinised by all assessors. Training was undertaken to ensure assessors from different institutions applied the same standards to study participants. Domains which assessors found ambiguous and difficult to interpret were addressed specifically to achieve consensus on the approach to marking. The methodology was designed so that DOPS was implemented according to ANZCA's intended use of the tool. Therefore trainees underwent a structured 'pre-procedure interview' to assess cognitive domains including clinical aspects of the block to be performed (Appendix). Assessors blinded to trainee identity and level of training scored videos offline. All six assessors scored all procedures (a fully crossed design). Assessors signed a privacy policy regarding the use and stewardship of videos thereby maintaining trainee and patient privacy.

Patient demographics [age, gender, body mass index (BMI), American Society of Anesthesiologists (ASA) physical status] nerve block performed, level of training (Basic Trainee Year 1, Basic Trainee Year 2, Advanced Trainee, Provisional Fellow) and the number of nerve blocks performed by the trainee prior to the study block were collected. Videos were presented to assessors in randomised order. Following a single viewing of videos assessors recorded the assessment and time taken to complete it to the online database. Data entry was monitored for completeness, accuracy and transcribing errors.

Psychometric analysis

DOPS includes 13 individual items: 11 9-point scale items and 2 dichotomous items. One of the 9-point scale items ('documentation/post-procedure management') could not be assessed

from the video and was not analysed. DOPS ‘total score’ was calculated as a percentage of the maximum score possible for all 9-point scale items excluding the ‘overall performance’ item. Items marked ‘unable to assess’ were excluded when calculating the ‘total score’. This reduced the maximum score possible so that percentages reflect trainee’s performance on domains which assessors felt they could assess. Inter-rater reliability of ‘total score’ and individual 9-point scale items were assessed with intraclass correlation coefficients (ICC) based on a two way random effects model. This model was chosen because both the blocks performed and the assessors were a random sample of all possible procedures and assessors respectively. Our ICC calculation takes into account absolute differences between scores, in addition to their correlation, and corresponds to ICC (A,1) according to McGraw and Wong.[15] A consistency model ICC (which only evaluates correlation and not actual agreement- ICC(C,1)[15]) was not used in this study because DOPS is a criterion-based assessment, therefore the actual value of the score is important. The ANZCA DOPS 9-point scale is grouped into three categories relating to the level of guidance required from the supervisor. Item ICCs were again calculated after the 9-point scale was recoded to a 3-point scale according to the aforementioned categories. The inter-rater reliability of the dichotomous item ‘does the trainee need to be re-assessed?’ was assessed by Cronbach’s alpha. The second dichotomous item (‘was procedure performed satisfactorily?) was not analysed because 'yes' was ticked by all assessors for nearly all blocks.

The internal consistency of DOPS was assessed in two ways: (i)Spearman rank correlation was calculated between ‘total score’ and response to the ‘overall performance’ item and (ii) ‘total score’ and ‘overall performance’ item response awarded to individuals who did, and did not, require re-assessment were compared using two-sample t and Mann-Whitney tests respectively.

Construct validity was evaluated by calculating correlation (Pearson's correlation co-efficient) between the 'total score' and the number of prior ultrasound-guided regional procedures performed by each trainee at the time of their videoed block. This correlation (Spearman rank correlation co-efficient) was also performed with the 'overall performance' item response. The trainees were divided into 2 equal groups – those who had performed fewer than the median number of blocks (36.5) and those who had performed more. The 'total score' and 'overall performance' item response awarded to individuals in these two groups were compared using two-sample t and Mann-Whitney tests respectively.

Feasibility was assessed by measuring time to complete assessments i.e. total time to complete assessment minus the actual video duration time. All statistical analyses were performed using Stata IC 12.1 (StataCorp, College Station, Texas). For all analyses, statistical significance was determined by a p value < 0.05.

Results

Thirty videos were viewed and assessed by offline assessors. Patients were 50% male, and had mean age 57 years (range 23 - 85), mean BMI 28.5 (range 20.2 – 40.3) and median ASA score II (range I – III). Procedures were performed by thirteen anaesthetic registrars (Basic Trainee Year 1 [n= 1], Basic Trainee Year 2 [n= 4], Advanced Trainee [n= 6], Provisional Fellow [n= 2]). Nerve blocks filmed included: femoral [n= 2 (7%)], saphenous [n= 5 (17%)], popliteal sciatic [n= 9 (30%)], interscalene [n= 5 (17%)], supraclavicular [n= 2 (7%)] and axillary (median nerve n= 3, radial nerve n= 3 and musculocutaneous n=1 [23%]). The number of nerve blocks previously performed by trainees ranged from 5 to 75 (mean= 39).Video duration ranged from 15 minutes 23 seconds to 36 minutes 30 seconds with a mean of 25 minutes 23 seconds. Six of 2340 (0.2%) database rows were not completed comprising estimates of ‘time to complete assessment’ from one assessor.

Table 1 provides summary statistics of the DOPS 9-point scale items and their inter-rater reliability. It also shows the ICCs when responses are recoded into a 3-point scale. Table 2 summarises the ‘total score’ values of each assessor and the inter-rater reliability of the 'total score'. The dichotomous item ‘does trainee need to be reassessed?’ had a Cronbach’s alpha of 0.85.

There was a correlation between ‘total score’ and the 'overall performance' item response – Spearman rho = 0.68, $p < 0.001$.

The mean ‘total score’ of procedures which assessors indicated that trainees did, and did not, need re-assessment were 62.3% and 76.6% respectively, $p < 0.001$. The median 'overall performance' item response for procedures which assessors indicated that trainees did, and did not, need re-assessment were 3 and 7 respectively, $p < 0.001$.

DOPS ‘total score’ correlated with the number of blocks previously performed by the trainee ($r = 0.51$, $p = 0.004$). This relationship is demonstrated in Figure 1. The ‘overall performance' item response also correlated with the number of blocks previously performed by the trainee (Spearman's rho = 0.45, $p = 0.01$). The median number of blocks trainees had previously performed was 36.5. Trainees who had performed fewer than 36.5 blocks had lower ‘total scores’ than trainees with more previous experience (mean 65.8% versus 73.0% respectively, $p = 0.01$). These trainees also achieved lower ‘overall performance' item responses than trainees with more previous experience (median 3.7 and 6 respectively, $p = 0.04$)

The mean time to assess the videos was 6 minutes 35 seconds (range: 2 minutes 37 seconds – 8 minutes 30 seconds).

Discussion

The inter-rater reliability (ICC) of individual 9-point scale items ranged between 0.10 and 0.49. When recoded to a 3-point scale, items had ICCs between 0.05 and 0.48. Technical ability demonstrated the highest ICC, 0.49 (95% C.I 0.33 – 0.66) and 'total score' was no more reliable than individual items (ICC = 0.25). Part of the rationale for using this particular DOPS is that it may be easier and more useful to assess supervision levels required, rather than actual competence. It was therefore significant that the reliability of the 3-point scale was not superior to the 9-point scale items. It is important to note studies evaluating reliability of assessment tools have used very different study designs and statistical techniques. For example, Sultan et. al. measured the inter-rater reliability of both a checklist tool (0.84) and global rating scale (0.79) for ultrasound-guided axillary block using two assessors with Cronbach's alpha.[16] It is important to appreciate that Cronbach's alpha measures correlation between assessors' scores, rather than actual agreement as was measured with the ICCs used in this current study. For example, adding a constant value (e.g. 2 points) to one assessor's scores would not alter the Cronbach's alpha, but would have a marked effect on the ICC we used.[15] This is because this ICC takes into account absolute differences between scores, as well as their correlation. In fact, had we used Cronbach's alpha to measure inter-rater reliability, the value for 'total score' and 'overall performance' item response would both have been 0.85 compared to our ICC values of less than 0.5. As a further example, Naik et. al. assessed the inter-rater reliability of a checklist (r = 0.85) and global rating scale (r = 0.74) for assessment of ultrasound-guided brachial plexus block with two assessors.[17] Their measure of inter-rater reliability was a Pearson's correlation coefficient, which again does not measure agreement between observers. In a review of anaesthetic procedural assessment methods, Bould outlines this limitation of using correlation statistics as a measure of inter-rater reliability.[5]

The aim of WBA is to “engender a coaching culture”[18] in training programs by drawing the focus toward “assessment for learning”, rather than “assessment of learning”.[19] Hence, DOPS may not need to possess a high degree of single-instance reliability compared with a tool used for high-stakes assessment. ICCs are measured from zero to 1, with higher values indicating greater correlation and/or agreement i.e. not correlation alone. Furthermore, the magnitude of an ICC as a measure of inter-rater reliability depends on the variability of scores i.e. the variability of the population to which it is applied.[15] For example, for a given degree of inter-rater variability, the ICC would be higher if the variability of the trainee scores were greater. ICCs have no accepted benchmarks describing ‘poor’, ‘moderate’ and ‘good’ reliability; in contrast to arbitrary levels set for kappa statistics [20] and Cronbach’s alpha [21].

Internal consistency of the DOPS tool was demonstrated by significant correlation between ‘total score’ and ‘overall performance’ item responses. Furthermore these measures were lower in those trainees who were deemed to require re-assessment indicates. Formal measures of internal consistency such as inter-item Cronbach’s alpha are probably not useful for a tool with items that assess a variety of cognitive and technical domains.

The construct validity of the ANZCA DOPS for assessment of ultrasound-guided regional anaesthesia was demonstrated by a relationship between ‘total score’ and ‘overall score’ item response and trainees’ level of experience. Furthermore, Figure 1 demonstrates that improvement in DOPS scores occurs within the first 30 blocks. This is consistent with the literature which indicates significant improvements in skills required for UGRA occur in the first 30 procedures.[22, 23]

The mean time to complete DOPS was approximately 20% of the video time. This is consistent with previous findings for video-assisted assessment and the use of DOPS in daily

practice.[24-27] Qualitative investigations using stakeholder feedback have produced mixed results of DOPS feasibility and acceptability.[26-29] The DOPS assessment required 6.5 minutes (on average), potentially supporting its feasibility in practice.

This study is the first to evaluate ANZCA DOPS' for reliability, validity and feasibility. This study has used a methodology which adhered closely to ANZCA's intended use of DOPS as a WBA. Our methodology has a number of important features: 1) assessing UGRA, now considered a core skill in anaesthesia; 2) the use of six assessors, each from a different institution; 3) the use of a variety of nerve blocks and 4) trainees at different levels of training and experience. We believe these are strengths because DOPS is intended to be used across Australia and New Zealand in a variety of institutions and for different procedures. Assessor training and the use of a pre-procedure interview were both guided by ANZCA's instructions for the use of DOPS.[30] Overall these results reflect an accurate estimate of the psychometric properties of ANZCA DOPS in the assessment of UGRA.

The potential limitations of this study are those inherent to all assessments using direct observation. These included the possibility of halo, contrast and assessor shift/drift.[31-33] Training in the use of assessment tools may not entirely attenuate the effect of these biases.[32] We sought to limit the impact of halo and contrast biases by presenting trainees to assessors in a randomised order. Assessor shift was unlikely given the short assessment period (six weeks) and assessor drift is unlikely to have occurred because assessors did not discuss marking after the training period.[31] Feedback is an essential part of DOPS's use but this study was not designed to evaluate this.[10]

This study is the first to evaluate the psychometric properties of the ANZCA DOPS assessment tool. Our methodology used a fully crossed design and took into account assessor agreement as well as correlation in the measurement of inter-rater reliability. The inter-rater

reliability (measured by ICC) of individual items varied from 0.10 to 0.49 and the inter-rater reliability of ‘total score’ was 0.25. The ANZCA DOPS demonstrated construct validity in the assessment of UGRA and is potentially feasible in daily practice.

Acknowledgements

The authors would like to thank Dr. Adam Dubrowski for his expert advice.

Competing interests

No external funding and no competing interests declared.

Supplementary Figure 1- ANZCA DOPS form[30]

See PDF attached: “ANZCA DOPS”

Appendix- Supervisor Instruction Sheet

Supervisor Instruction Sheet: Supervisor Prompts

Do not rephrase. Read out questions (in italics) in sequential order.

Pre-procedure interview (away from patient)

1. *What is the **name of the block** and what is the **indication** for the block?*

2. *What are the most important **anatomical landmarks**?*

   (Provide clues if necessary: Nerves, vessels, muscles, fascial planes if relevant)

3. *What **sensory** and **motor** blockade will the block produce?*

4. *Assume I am the patient. Please give me **informed consent** for this block procedure.*

   (Provide clues if necessary: Side effects, risks and complications)

5. *What **monitoring** are you going to use during the block?*

6. *What **probe** are you going to use?*

(If trainee states anything other than high or intermediate frequency/ linear probe, ask for further explanation).

7. *Patient is …kg:*
   What are you going to inject *and what* ***volume*** *and* ***concentration****? What* ***needle type****?*

8. Post-procedure management:

   How are you going to test the success of the block?

   If the block is ***unsuccessful*** *(or incomplete) what are you going to do (assuming you still plan to proceed for surgery)?*

   If the block is ***successful*** *what instructions are you going to give the patient and the nurse looking after the patient, after the operation?*

Instructions for the anaesthetic nurse (show the nurse but not the trainee)

9. Do not set up the room, position the patient or initiate the time-out without being directed

10. Pretend to be an "untrained assistant"

11. Interrupt at any time if safety is at risk, (e.g. step-in if a time-out is not done, but give the trainee a chance to initiate this).

Scanning

12. *Show me the most important anatomical landmarks and tell me their names.*

13. *Are there any other important anatomical features?*

    (Provide clues if necessary: Nerves, vessels, muscles, fascial planes if relevant)

14. *Show me the* ***best image*** *of the nerve.*

    (Provide clues if necessary: Is there anything you can do to improve the image?)

15. *Where are you aiming to position your needle tip in relation to the nerve?*

    (Use O'Clock descriptions if helpful)

16. *Where would you like to see the* ***injectate*** *spread?*

    (Use O'Clock descriptions if helpful)

Needle proficiency

17. Remind the trainee to make comments about the **needle tip position** throughout the procedure. *"Where is the needle tip?"*

18. **Problem solving**. If difficult ask, *"do you know any tricks to help make the needle more visible?"*

Injection

Immediately prior to injection

19. *Before you inject,* ***what instructions*** *are you going to give me regarding injection, assuming I am an untrained assistant?*

Immediately after injection

20. ***Where can you see the injectate*** *going and is it going where you want it to?*

1 Wallenburg I, Helderman J-K, de Bont A, Scheele F,Meurs P. Negotiating Authority: A Comparative Study of Reform in Medical Training Regimes. *Journal of Health Politics, Policy and Law* 2012 June 1, 2012; **37**(3): 439-467
2 Drazen JM,Epstein AM. Rethinking Medical Training — The Critical Work Ahead. *New England Journal of Medicine* 2002 **347**(16): 1271-1272
3 Ludmerer KM,Johns M. E. Reforming graduate medical education. *JAMA* 2005 **294**(9): 1083-1087
4 Fleet LJ, Kirby F, Cutler S, Dunikowski L, Nasmith L,Shaughnessy R. Continuing professional development and social accountability: A review of the literature. *Journal of Interprofessional Care* 2008 **22**(s1): 15-29
5 Bould MD, Crabtree NA,Naik VN. Assessment of procedural skills in anaesthesia. *British Journal of Anaesthesia* 2009 August 30 **103**(4): 472-483
6 Darzi A, Smith S,Taffinder N. Assessing operative skill. *BMJ* 1999 1999-04-03; **318**(7188): 887-888
7 Hays RB, Davies HA, Beard JD, et al. Selecting performance assessment methods for experienced physicians. *Medical Education* 2002 **36**(10): 910-917
8 Fromme HB, Karani R,Downing SM. Direct Observation in Medical Education: A Review of the Literature and Evidence for Validity. *Mount Sinai Journal of Medicine: A Journal of Translational and Personalized Medicine* 2009 **76**(4): 365-371
9 Meyers FJ, Weinberger SE, Fitzgibbons JP, et al. Redesigning Residency Training in Internal Medicine: The Consensus Report of the Alliance for Academic Internal Medicine Education Redesign Task Force. *Academic Medicine* 2007 **82**(12): 1211-1219
10 Wragg A, Wade W, Fuller G, Cowan G,Mills P. Assessing the performance of specialist registrars. *Clinical Medicine* 2003 2003 March/April **3**(2): 131-134
11 Anaesthesia training program: Curriculum. Regional and local anaesthesia. Australian and New Zealand College of Anaesthetists, Melbourne 2012. http://www.anzca.edu.au/training/2013-training-program/pdfs/anaesthesia-training-program-curriculum.pdf- (Accessed 2013 March 18)
12 ANZCA. ANZCA Handbook for Training and Accreditation. ANZCA, Melbourne 2012. http://www.anzca.edu.au/training/2013-training-program/pdfs/training-accreditation-handbook- (Accessed 2013 March 18)
13 GMC. Standards for curricula and assessment systems. Delivery. General Medical Council, London 2010. http://www.gmc-uk.org/Standards_for_Curricula__Assessment_Systems.pdf_31300458.pdf- (Accessed 2013 June 14)
14 Van Der Vleuten CP. The assessment of professional competence: Developments, research and practical implications. *Advances in Health Sciences Education* 1996 Jan; **1**(1): 41-67
15 McGraw KO,Wong SP. Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods* 1996 **1**(1): 30-46
16 Sultan SF, Iohom G, Saunders J,Shorten G. A clinical assessment tool for ultrasound-guided axillary brachial plexus block. *Acta Anaesthesiologica Scandinavica* 2012 **56**(5): 616-623
17 Naik VN, Perlas A, Chandra DB, Chung DY,Chan VWS. An Assessment Tool for Brachial Plexus Regional Anesthesia Performance: Establishing Construct Validity and Reliability. *Regional Anesthesia and Pain Medicine* 2007 **32**(1): 41-45
18 Shaw M. Road-testing workplace-based assesments Australian and New Zealand College of Anaesthetists, 2011 September 2011 http://www.anzca.edu.au/training/2013-training-program/pdfs/Road-testing%20workplace-based%20assessment.pdf (Accessed 18.3.13)
19 Schuwirth LWT,Van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher* 2011 **33**(6): 478-485
20 Landis JR,Koch GG. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics* 1977 **33**(2): 363-374
21 Gliem J,Gliem R. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Midwest Research-to-Practice Conference in Adult, Continuing, and

Community Education, 2003 (Conference Proceedings). http://hdl.handle.net/1805/344 (Accessed 7/2/2013)
22 Barrington MJ, Wong DM, Slater B, Ivanusic JJ,Ovens M. Ultrasound-Guided Regional Anesthesia: How Much Practice Do Novices Require Before Achieving Competency in Ultrasound Needle Visualization Using a Cadaver Model. *Regional Anesthesia & Pain Medicine May/June* 2012 **37**(3): 334-339
23 de Oliveira Filho GR. The Construction of Learning Curves for Basic Skills in Anesthetic Procedures: An Application for the Cumulative Sum Method. *Anesthesia & Analgesia* 2002 August 1, 2002; **95**(2): 411-416
24 Driscoll PJ, Paisley AM,Paterson-Brown S. Video assessment of basic surgical trainees' operative skills. *The American Journal of Surgery* 2008 **196**(2): 265-272
25 van Hove PD, Tuijthof GJM, Verdaasdonk EGG, Stassen LPS,Dankelman J. Objective assessment of technical surgical skills. *British Journal of Surgery* 2010 **97**(7): 972-987
26 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): 364-373
27 Bindal N, Goodyear H, Bindal T,Wall D. DOPS assessment: a study to evaluate the experience and opinions of trainees and assessors. *Medical Teacher* 2013 Jun; **35**(6): e1230-1234
28 Cohen SN, Farrant PBJ,Taibjee SM. Assessing the assessments: U.K. dermatology trainees' views of the workplace assessment tools. *British Journal of Dermatology* 2009 **161**(1): 34-39
29 Finall A. Trainers' perceptions of the direct observation of practical skills assessment in histopathology training: a qualitative pilot study. *Journal of Clinical Pathology* 2012 **65**(6): 538-540
30 DOPS WBA Form. Australian and New Zealand College of Anaesthetists 2012. http://www.anzca.edu.au/training/2013-training-program/pdfs/direct-obervation-of-procedural-skills.pdf (Accessed 2nd Feburary 2013).
31 Harris FC,Lahey BB. Recording system bias in direct observational methodology: A review and critical analysis of factors causing inaccurate coding behavior. *Clinical Psychology Review* 1982 **2**(4): 539-556
32 Margolis MJ, Clauser BE, Cuddy MM, et al. Use of the Mini-Clinical Evaluation Exercise to Rate Examinee Performance on a Multiple-Station Clinical Skills Examination: A Validity Study. *Academic Medicine RIME: Proceedings of the Forty-fifth Annual Conference October 29-November 1, 2006* 2006 October; **81**(10): S56-S60
33 Yeates P, O'Neill P, Mann K,Eva K W. Effect of exposure to good vs poor medical trainee performance on attending physician ratings of subsequent performances. *JAMA* 2012 **308**(21): 2226-2232

Table 1 Summary statistics of DOPS item scores and inter-rater reliability of DOPS individual items.

Values are median (IQR) and Intraclass correlation coefficients (95% C.I)

	9-point scale			3 category scale
DOPS item	Median score (IQR)*	Range*	ICC (95% C.I)	ICC (95% C.I)
Clinical knowledge	7 (5 – 8)	1 – 9	0.34 (0.17 – 0.53)	0.24 (0.11 – 0.43)
Consent	7 (6 – 8)	3 – 9	0.21 (0.76 – 0.39)	0.23 (0.10 – 0.42)
Preparation	7 (6 – 8)	2 – 9	0.20 (0.07 – 0.39)	0.19 (0.06 – 0.36)
Vigilance†	7 (6 – 8)	1 – 9	0.10 (0.01 – 0.27)	0.06 (-0.02 – 0.21)
Infection control	8 (7 – 8)	1 – 9	0.10 (0.02 – 0.24)	0.14 (0.04 – 0.29)
Technical ability	6 (4 – 7)	1 – 9	0.49 (0.33 – 0.66)	0.48 (0.33 – 0.66)
Patient interaction‡	6 (5 – 7)	1 – 9	0.35 (0.19 – 0.55)	0.30 (0.15 – 0.49)
Insight§	7 (5 – 7)	2 – 9	-	-
Team interaction	7 (6 – 8)	2 – 9	0.15 (0.05 – 0.30)	0.05 (-0.01 – 0.16)
Overall performance for this procedure	4 (3 – 7)	1 – 9	0.32 (0.14 – 0.52)	0.27 (0.12 – 0.47)

DOPS= Direct observation of procedural skills form; IQR= Interquartile range; ICC= Intraclass correlation coefficient. DOPS item “Documentation/ post-procedure management” not analysed because of poor response rate.

* Median (Interquartile range) and range of all six assessors’ scores

† Vigilance was unable to be assessed in 6.7% of assessments.

‡ Patient interaction was unable to be assessed in 1.7% of assessments

§ ICC not calculated for Insight as response rate was too low (was unable to be assessed in 24% of assessments).

Table 2 Summary of ‘total score’ provided by each assessor and overall inter-rater reliability. Values given as range, median (IQR) and intraclass correlation coefficients.

Assessor		1	2	3	4	5	6	Overall
‘Total score’ (% of possible mark)	Range	57 - 96	50 - 89	58 - 85	42 - 89	50 - 97	29 - 72	29 – 97
	Median (IQR)	79 [74 – 86]	75 [64 – 81]	72 [69 – 75]	65 [54 – 74]	80 [67 – 84]	51 [42 – 59]	72 [60 – 80]
	ICC*	0.25						
	95% C.I*	0.08 – 0.46						

ICC= Intraclass correlation coefficient; IQR= Interquartile range.

*ICC and 95% C.I for ‘total score’ inter-rater reliability

Figure 1 Graph of trainee's 'total score' versus the number of blocks performed prior to study block. 'Linear fit' (dotted line) is the line of best fit as determined by linear regression. 'Line of best fit' (solid line) is determined using locally weighted regression and assumes no specific type of relationship between the variables.

See TIFF file attached: Figure 1