Assessment of procedural skills in anaesthesia

M. D. Bould1*, N. A. Crabtree2 and V. N. Naik3

1The Hospital for Sick Children, University of Toronto, 555 University Avenue, Canada M5B 1W5. 2Oxford Radcliffe Hospitals NHS Trust, Headley Way, Oxford OX3 9DU, UK. 3The Ottawa Hospital, University of Ottawa, 1053 Carling Avenue, Ottawa, Ontario, K1Y 4E9, Canada

*Corresponding author. E-mail: dylan.bould@utoronto.ca

A key aspect of the practice of anaesthesia is the ability to perform practical procedures efficiently and safely. Decreased working hours during training, an increasing focus on patient safety, and greater accountability have resulted in a paradigm shift in medical education. The resulting international trend towards competency-based training demands robust methods of evaluation of all domains of learning. The assessment of procedural skills in anaesthesia is poor compared with other domains of learning and has fallen behind surgical fields. Logbooks and procedure lists are best suited to providing information regarding likely opportunities within training programmes. Retrospective global scoring and direct observation without specific criteria are unreliable. The current best evidence for a gold standard for assessment of procedural skills in anaesthesia consists of a combination of previously validated checklists and global rating scales, used prospectively by a trained observer, for a procedure performed in an actual patient. Future research should include core assessment parameters to ensure methodological rigor and facilitate robust comparisons with other studies: (i) reliability, (ii) validity, (iii) feasibility, (iv) cost-effectiveness, and (v) comprehensiveness with varying levels of difficulty. Simulation may become a key part of the future of formative and summative skills assessment in anaesthesia; however, research is required to develop and test simulators that are realistic enough to be suitable for use in high-stakes evaluation.

Br J Anaesth 2009; 103: 472–83

Keywords: education; risk; safety
assessment of procedural skills has been pioneered in surgery.\textsuperscript{59} We will therefore also discuss innovations in the surgical field that may be useful for anaesthesia in the future. This review will discuss the acquisition of expertise in procedural skills, general principles about the educational theory behind assessment, examine literature regarding the different techniques that can be used for the assessment of procedural skills in anaesthesia and make relevant extrapolations from literature in other fields of medicine. Finally, we will discuss the context of procedural skills assessment, including the use of simulation.

The acquisition of expertise in procedural skills

There are three stages in the acquisition of procedural skills: cognition, integration, and automation.\textsuperscript{52} Cognition includes developing an understanding of the task, and perceptual awareness. It is assisted by a clear description and demonstrations of the task. In the integration stage, the knowledge from the cognition phase is incorporated into the learning of the motor skills for that task. Ultimately, the task becomes automatic and even subconscious. For an expert, it may be difficult to break a task down into component parts in order to teach a novice.

The acquisition of competence or expertise in a procedural skill requires experience through a variable number of attempts depending on the skill, the quality of teaching, and the aptitude of the individual. In anaesthesia, skills are typically initially learned on relatively straightforward cases. When competence is achieved in those situations, trainees are exposed to a broader range of normal and pathological variations required for the development of true expertise. Retention of motor skills seems to be most dependent on the degree to which the skill was perfected.\textsuperscript{37}

Which skills should be assessed?

A full discussion of which skills should be assessed is beyond the scope of this review. One viewpoint is that technical skills should be assessed if they are either commonly performed or potentially life saving.\textsuperscript{93} However, determining the core skills for life-saving manoeuvres can be controversial, as demonstrated by a recent debate on determining the core skills for life-saving manoeuvres can be perfected.\textsuperscript{37}

Reliability

Reliability refers to the reproducibility of a test. In education, this may refer to inter-rater agreement or test–retest agreement. These agreements are often described as ‘external reliability’. There are different statistical analyses that can be used to describe reliability:

(1) Pearson’s product-moment correlation coefficient has been used to describe inter-rater agreement. An $r >0.75$ indicates excellent agreement.\textsuperscript{64} However, this method has been criticized as not accounting for bias.\textsuperscript{43} For instance, if one examiner consistently uses the low marks in a scale and another consistently uses the higher marks, they could still have a high correlation coefficient if they rank subjects similarly. This bias would be particularly significant if the assessment tool has a set pass mark.

(2) Intraclass correlation coefficient (ICC). The ICC is also used to describe inter-rater agreement. It accounts for the agreement that would be seen by chance and is
defined as the ratio of variance between subjects due to error variance. An ICC of 0.8 means that 80% of the variance among scores can be attributed to true variance among subjects. Cohen's $\kappa$ coefficient is a type of ICC that can only be used when there are two raters. A $\kappa > 0.80$ has been described as indicating near-perfect agreement; $0.61-0.80$, substantial agreement; $0.41-0.60$, moderate agreement; $0.21-0.40$, fair agreement; $0.00-0.20$, slight agreement; and $<0.00$, poor agreement. However, it should be noted that this categorization is not universally accepted, and in practice, the degree of acceptable agreement depends on the circumstances. For instance, a ‘high stakes’ licensing exam requires a particularly reliable assessment tool.

Internal consistency is sometimes described as ‘internal’ reliability, when compared with the ‘external’ reliability of inter-rater or test–retest agreement. Internal consistency is not whether a subject performs similarly in each part of a test, but whether different subjects tend to do well or badly on the same parts of the test. Cronbach’s $\alpha$ gives a value from 0.0 to 1.0. By convention 0.0–0.5 can be regarded as imprecise, 0.5–0.8 moderately reliable, and 0.8–1.0 can be used with confidence for high-stakes purposes such as certification, although again these cut-offs are arbitrary. Internal consistency is a commonly used measure of reliability as it describes the reproducibility of a test. However, the consequences of poor internal reliability are less problematic than poor external reliability. For example, a test where different candidates fail on different questions could be due to differences in clinical experience or teaching, whereas a test with poor inter-rater reliability could be considered intrinsically unfair to the subjects.

**Validity**

Validity describes whether the test is measuring what it sets out to measure.

(1) Face validity refers to a general impression of whether the evaluation seems appropriate. For example, evaluating performance of epidural anaesthesia on a model consisting of a banana has little face validity compared with direct observation of the placement of an epidural in a patient. Face validity is perhaps best assessed by expert opinion, although good face validity as judged by the subjects of the assessment improves ‘buy-in’ of the evaluation.

---

### Table 1: Reliability, validity, and feasibility of techniques to assess technical skills

<table>
<thead>
<tr>
<th>Test</th>
<th>Reliability</th>
<th>Validity</th>
<th>Feasibility</th>
<th>Comprehensiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychometric testing</td>
<td>Automated and objective</td>
<td>Limited correlation with early performance in fibreoptic endoscopy</td>
<td>Relatively expensive, requires a degree of expertise to be able to collect, process and interpret the results</td>
<td>Not comprehensive</td>
</tr>
<tr>
<td>Procedural lists</td>
<td>Self-reported and subject to omissions and inaccuracies</td>
<td>No guarantee that procedures were performed correctly</td>
<td>The easiest form of evaluation</td>
<td>Not comprehensive</td>
</tr>
<tr>
<td>Procedural lists with</td>
<td>Self-reported and subject to omissions and inaccuracies</td>
<td>No guarantee that procedures were performed correctly</td>
<td>Very easy to use</td>
<td>Not comprehensive</td>
</tr>
<tr>
<td>prescribed minimum number</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cusum</td>
<td>If self-reported is subject to omissions and inaccuracies</td>
<td>If self-reported, there is no guarantee that procedures were performed correctly</td>
<td>Requires esoteric statistical analysis, usually depends on self-reporting</td>
<td>Depends on how success and failure is defined and identified</td>
</tr>
<tr>
<td>Direct observation</td>
<td>Poor reliability</td>
<td>Has face validity</td>
<td>Easy to use, requires some training for optimal reliability</td>
<td>Potentially comprehensive: depends on content of checklist</td>
</tr>
<tr>
<td>without criteria</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Checklists</td>
<td>Excellent reliability in trained observers, especially within the context of an educational study</td>
<td>Construct validity established in anaesthesiology for lumbar epidurals and interscalene blocks. Intrinsic content validity</td>
<td>Easy to use, requires some training for optimal reliability</td>
<td>Potentially comprehensive: depends on content of global ratings scale</td>
</tr>
<tr>
<td>Global rating scales</td>
<td>Excellent reliability in trained observers, especially within the context of an educational study</td>
<td>Construct validity established in anaesthesiology for lumbar epidurals and interscalene blocks. Intrinsic content validity</td>
<td>Easy to use, requires some training for optimal reliability</td>
<td>Potentially comprehensive: depends on content of global ratings scale</td>
</tr>
<tr>
<td>Motion analysis</td>
<td>Objective with a numerical output and so not subject to differences in the opinion of the observer</td>
<td>Construct validity established for open general surgery, laparoscopic surgery, and microsurgery but as yet untested in anaesthesia</td>
<td>Relatively expensive, requires a degree of expertise to be able to collect, process, and interpret the results</td>
<td>Not comprehensive: only evaluates motor skill</td>
</tr>
<tr>
<td>Simulation</td>
<td>Depends on assessment tool used with the simulation</td>
<td>Highly variable degrees of face validity between simulations. Otherwise, validity is largely unknown</td>
<td>Cost is highly variable depending on the type and complexity of the simulation</td>
<td>Depends on assessment tool and the nature of the simulation</td>
</tr>
<tr>
<td>Multi-station bench testing</td>
<td>Excellent reliability in trained observers, especially within the context of an educational study</td>
<td>Construct validity established for open general surgery, laparoscopic surgery but as yet no comparable assessment tool in anaesthesia</td>
<td>Relatively expensive, time-consuming for faculty, requires co-ordination at a central site</td>
<td>Potentially comprehensive</td>
</tr>
<tr>
<td>(e.g. OSATS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(2) Content validity refers to whether an assessment tests the content either of what was being taught or appropriate content as defined by a group of experts.
(3) Concurrent validity establishes validity based on agreement with another established valid measure.
(4) Construct validity is used when there is no established gold standard for comparison. A construct is a concept that is to some extent abstract. For instance, although we can all agree that there is such a thing as ‘expertise’, it is more difficult to precisely define or test that construct. Instead, we can test a surrogate outcome, such as experience, that is easier to quantify and that we expect to be associated with the construct of expertise. An educational evaluation is therefore considered valid if it can differentiate between groups with different levels of experience and is increasingly valid if the groups it can distinguish between are more similar in experience. Reznick comments that ‘validity cannot be proven in any one experiment. Rather, over time and experimentation one accrues evidence for the validity of a test’.68 This is particularly true for construct validity, which only examines a surrogate measure.
(5) Predictive validity is the ability of a test to predict something that happens after the test such as a clinical outcome or the future test results of a subject. Although such outcome data are the most useful demonstration of validity, it is generally the most difficult to establish.

Techniques for the assessment of procedural skills

Psychometric and aptitude testing
Psychometric testing has been found to be of limited value in predicting subsequent procedural performance in the surgical field. A statistically significant but modest correlation was found between the performance of a Z-plasty procedure and scores in tests that assess the ability to rotate 2D and 3D figures mentally, but not with less complex tests that assess the recognition of simple shapes.92 Laparoscopic surgery requires surgeons to infer the shape of 3D structures from 2D screens. The Pictorial Surface Orientation (PicSOr) is a computer-based test of depth perception.31 Gallagher and colleagues31 compared PicSOr performance to simulated laparoscopic cutting tasks in both novices and expert laparoscopic surgeons. There was a modest correlation between performance in PicSOr and laparoscopic performance.

The MICROcomputerised Personnel Aptitude Tester (MICROPAT) measures psychomotor ability and has been used in anaesthesia to compare performance in adaptive pursuit tracking tasks with subsequent performance in fibreoptic nasotracheal endoscopy. Pursuit tracking was correlated with faster times to completion of nasotracheal endoscopy, accounting for approximately one-third of the ability early in the learning curve.17 The MICROPAT has also been investigated as a method of predicting obstetric epidural failure rates but was not correlated with failure rates for either the first 25 or the first 50 epidurals suggesting a limited application for this evaluation.16

As psychometric tests have only been shown to have moderate correlations with performance in the early stages of technical skill acquisition, it remains to be seen if they have a role in medical recruitment or selection for any specialty. Indeed, evidence from fields of expertise outside medicine suggests that large numbers of hours of deliberate practice are more important to the acquisition of motor skills than innate ability.21 In the context of procedural skills in anaesthesia, psychometric testing is currently essentially a research tool and largely unproven.

Procedure lists
Assessment of technical skill has historically been by a combination of a subjective impression from an educational supervisor and logbooks or procedural lists.96 Self-reported procedure lists are a common form of assessment of technical skills mainly because of high feasibility. Although a certain number of procedures performed are clearly necessary to provide the opportunity to progress through the stages of acquisition of technical skills, the actual number is highly variable between individuals.15 20 There are clear limitations in the value of procedure lists, especially if used for summative assessment: there is no guarantee that a task was performed correctly and trainee anaesthetists can consistently repeat mistakes, despite considering their own performance to be acceptable. Performing a procedure badly a large number of times is of little educational value and also puts patients at risk.93 In conclusion, procedural lists are most useful for assessing opportunities provided by training posts and to guide programme directors rather than to assess individuals.8

Cumulative sum analysis
Cumulative sum (Cusum) analysis was originally developed in industry as a method of quality control. It is a statistical method that looks at the outcome rather than the process of performing procedural skills. Cusum plots a graph of the subject’s performance over time based on predetermined criteria for success or failure. A value for Cusum is plotted on the y-axis and the number of attempts on the x-axis (Appendix). Failures result in a move up (and successes down) the y-axis as the subject progresses through increasing numbers of attempts. When the Cusum score decreases below a level based on a predetermined ‘acceptable failure rate’, the subject can be considered to be competent with statistical significance. The distance that a trainee is above a predetermined line is an indication of how far they are from achieving competency as defined
by Cusum. Authors have also described other useful endpoints such as a change in the curve that denotes either an improvement or worsening in performance.

The Cusum analysis is an effective objective tool to define learning curves for technical skills. Learning curves can be constructed for individuals or summed to provide learning curves for a population. Cusum can be used to provide an estimate of the number of cases that are required to achieve competency and demonstrates a wide variety in that number between individuals. For example, subjects required between nine and 88 attempts to become competent at tracheal intubation. It can be used to identify when a change in the training process is indicated, and poor technique may be corrected before demoralization. Cusum has been used to alter the schedule of training rotations, change curricula, and initiate mentoring programmes. As such, it can be used not only to assess the competency of individuals in procedural skills but also as an assessment of a training programme and its ability to teach those skills. Another potential application is as a continuous audit of quality of practice for experienced clinicians, although it is more commonly used this way in surgery where complications that constitute negative endpoints are more common.

A potential disadvantage of the Cusum method is that it often relies on self-reporting, which may be inaccurate. Direct observation of all procedures by trainees is unlikely to be possible for most procedures, as a considerable amount of work may be performed while on call. Cusum is only as objective as the pre-defined success/failure endpoint and these definitions may vary widely. Acceptable rates of success can be determined by institutional rates or expert consensus and also depends on the definition of success. As training in anaesthesia is based on the principle of gradually increasing responsibility and reducing supervision, an increase in responsibility may result in more difficult cases and therefore result in a deterioration of the Cusum curve, despite no deterioration in skill. The issue of accounting for the difficulty level posed to the subject can also be problematic for other methods of assessing technical skill and is further discussed in the simulation section.

A final disadvantage of Cusum is that a great number of attempts may be necessary to prove statistical significance and this could be unfair to trainees if their progression through a training rotation depends on Cusum defined competency.

**Direct observation without criteria**

Direct observation by a consultant is traditionally used to assess procedural skills and is feasible in anaesthesia because of the high degree of supervision in cases performed by trainees. Despite being feasible, assessments without specific criteria result in poor reliability and validity. To overcome these problems, direct observation with specific criteria has been developed.

**Direct observation with criteria**

**Checklists**

Binary content checklists can be used as a way of grading performance during direct observation. Checklists break a task down into its component parts and assign a dichotomous pass/fail outcome to each point. A new checklist needs to be designed and validated for each procedural skill that is to be assessed. Checklists can be constructed by surveying experts, although a different group of experts may not agree on each point of the checklist. For instance, different groups have published checklists for epidural anaesthesia with 14, 27, and 61 items on the list. A systematic review and content analysis of checklists for procedural skills assessment can be found elsewhere. Checklists have also been designed with outcomes of 'not performed', 'performed poorly', and 'performed well' rather than a binary pass–fail outcome to allow them to become more qualitative at the cost of a loss in objectivity. A potential problem with checklists is that if all stages are weighted equally regardless of clinical importance, then a trainee may be able to obtain a high score, despite omitting important stages. To prevent this, certain stages can be marked as resulting in an automatic fail if not completed or an overall pass/fail option can be added to the scoring system.

Checklists have been found to have excellent reliability in the assessment of epidural anaesthesia and good reliability for the assessment of interscalene brachial plexus blocks when used by trained assessors. An advantage of checklists is that they have intrinsic content validity, if they are constructed well, for instance, ensuring that the checklist examines what is taught at that centre. Construct validity has been established for checklists in the assessment of epidural anaesthesia and interscalene brachial plexus blocks. Predictive validity has not been studied for checklists in anaesthesia.

**Global rating scales**

Global rating scales (GRSs) differ from checklists, in that they use a Likert scale rather than a dichotomous outcome. As the GRS has a gradation of response in each category, it is less objective than a checklist, although this allows the assessment to be more qualitative. GRS can be used to assess many different skills and they are the most objective way that aspects of performance such as professionalism and interpersonal skills can be assessed. When used to assess procedural skill, a GRS may either describe an overall impression of the quality of performance or there may be a Likert scale for a number of different domains within an overall performance. GRS can be used prospectively or retrospectively, although, as in other forms of assessment, there is evidence that reliability...
is poor if used retrospectively. Potential pitfalls with GRS include the ‘halo effect’, when good or bad performance in one domain unduly influences the grading of performance in other domains. This may be partly due to lack of training of assessors and can make a GRS seem falsely internally consistent. Another common problem with GRS is self-imposed scale limitation. Assessors commonly restrict themselves to the high end of the scale: in one study, 95.6% of scores on a nine-point scale were between 6 and 9. This may also be because of either lack of training of assessors or because assessors are unwilling to fail a trainee knowing the potentially serious consequences. An alternative explanation for scale limitation is that most trainees are of a high standard. However, to be useful, GRS must be designed to be able to differentiate grades of quality beyond distinguishing between outstanding or failing trainees.

A GRS developed for the assessment of procedural skills in surgery at the University of Toronto (Table 2) has been repeatedly found to have construct validity for the assessment of procedural skills in both surgery and anaesthesia, differentiating between junior and senior trainees performing an interscalene block, and to discriminate between various levels of experience at performing epidural anaesthesia. It has also been found to have good reliability for the assessment of orotracheal fiberoptic intubation, epidural anaesthesia, and interscalene brachial plexus blocks.

An advantage of GRS is that they are not confined to one procedure but can be used for different procedural skills. However, some domains may be particularly useful for certain kinds of procedure, for instance, the domain ‘depth perception’ has been added for laparoscopy but is unlikely to be useful in anaesthesia while ‘autonomy’ is likely to be useful for procedural skills assessment in any speciality.

GRSs are currently being used for in-training assessment of procedural skills in the UK Foundation Programme that covers the first 2 postgraduate years in all specialities. One mandatory competence assessment tool is the Direct Observation of Procedural Skills (DOPS), a specific six-point, 11 domain GRS used to assess performance in procedural skills (Table 3). It is notable that DOPS...
focuses on the context of the procedural skill: nine of the domains describe pre- and post-procedure care and non-technical skills. The actual assessment of the procedural skill is limited to a single domain. DOPS was developed by the Royal College of Physicians UK and is currently in the process of being studied as a pilot system. Foundation year trainees need to undertake 6 DOPS each year from an approved list. The trainee chooses the timing, the procedure, and the assessor who may be a more senior doctor or a nurse but is expected to have had some training in the use of DOPS.

McKinley and colleagues argue for a holistic evaluation of procedural skills for summative purposes. Their Leicester Clinical Procedure Assessment Tool (LCAT) was developed in response to a systematic review of checklists and GRS for the evaluation of procedural skills that found that teamwork competencies and humanistic competencies such as safety and infection prevention were omitted from the majority of assessment tools. The LCAT has been shown to be reliable and has thoroughly demonstrated content and face validity, but construct and predictive validity have not yet been investigated.

It is not clear whether construct validity as tested within an educational research trial is necessarily generalizable to ‘real world’ practice using in-training assessments by untrained raters. Potential impediments to similar degrees of reliability in the ‘real world’ to that achieved in research with direct observation of a procedure on actual patients include an imprecise GRS, patient variability resulting in heterogeneous levels of difficulty, lack of training of the assessor resulting in differing levels of expectation from staff and the degree to which the trainee is acting independently.

Comparisons between GRSs and checklists
Both checklists and GRSs have been found to have good reliability. Checklists have been challenged as being able to distinguish novice and expert performance but failing to differentiate between high levels of performance, rewarding thoroughness rather than expertise. This may be particularly true for non-technical skills: Hodges and colleagues compared checklists and GRSs in psychiatry OSCE stations and found that experts scored better than trainees or medical students when assessed with GRS but worse when assessed with checklists. The authors concluded that checklists might penalize experts who take shortcuts that they have learned as part of their expertise and that ‘an instrument that is valid at one level of training may not be valid at another’.

Similarly, for procedural skills assessment, a comparison of checklists and GRSs in the performance of simulated surgical procedures by trainees found that GRS had better construct validity than the checklist, although both instruments demonstrated construct validity and good reliability.

Other authors have suggested that checklists may be better suited for assessing procedural skills than other domains of learning as procedural skills tend to be sequential and predictable. It has also been suggested that procedure-specific scales may provide an additional degree of formative feedback to generic scales. A good example is in ultrasound-guided regional anaesthesia, where radiological visualization of the nerve is a key skill that is not accounted for by generic scales. It is notable that GRS generally gather different information than a checklist. Friedman and colleagues have demonstrated the value of a checklist in identifying poor aseptic technique, despite good performance in other aspects of procedural skills. This suggests that a combination of a checklist and GRS may be advantageous when a comprehensive evaluation is required, for instance, when testing an intervention in education research.

A dilemma for both GRSs and checklists is determining cut-off scores for what can be considered competent or not competent if used for summative evaluation. One solution is to examine a population of expert anaesthetists and define proficiency from their scores. Using a normative marking scheme with a fixed proportion of subjects passing and failing has the limitation that it would fail to identify either a good or bad cohort of trainees.

Other instruments
Global Operative Assessment of Laparoscopic Skills (GOALS) is an assessment tool that combines both a GRS and a checklist with visual analogue scales (VAS). The two 10 cm VAS are anchored at each end with specific descriptors and refer to (i) overall competence and (ii) the observed difficulty of the procedure. GOALS was found to be more reliable than a GRS and checklist alone with suitable inter-rater reliability for a high-stakes examination. The VAS for competence was found to have construct validity in the assessment of performance of laparoscopic cholecystectomy where a checklist did not. The use of a VAS for difficulty when directly observing procedural skills has not yet been explored in anaesthesia and future research in this area is warranted. Another alternative to checklists and GRSs is to measure the number of predetermined errors.

Motion analysis
The Imperial College Surgical Assessment Device (ICSAD) is a motion analysis device originally designed for the investigation of hand movements in surgeons. It provides an objective measure of technical ability that has been validated in various surgical fields and has begun to be used in anaesthesia. It uses an electromagnetic tracking system (Isotrak II; Polhemus Inc., Colchester, VT, USA) consisting of an electromagnetic field generator and two 10 mm sensors that are attached to the dorsum of each hand. Robotic Video and Motion Analysis Software retrieves time-stamped Cartesian coordinates and defines hand movements by changes in velocity. It processes this
information to produce values for total distance moved by each hand, number of movements, total time, and average hand velocity. A Gaussian filter is used to eliminate background noise, so that only meaningful actions register as movements: the size of movement that designates a movement is adjustable; however, after this point, the ICSAD is entirely objective.

Different aspects of economy of movement have been validated as assessment tools for different procedures. Time taken and reduced total movements have been demonstrated to be associated with expertise in open\textsuperscript{51} \textsuperscript{9} laparoscopic\textsuperscript{83} and micro-surgery.\textsuperscript{75} Reduced number of movements has also been found to be correlated with less anastomotic leakage in a simulated arterial graft model providing some evidence of predictive validity.\textsuperscript{18} Reduced total path length is associated with expertise in laparoscopic\textsuperscript{59} \textsuperscript{81} \textsuperscript{83} and micro-surgery\textsuperscript{75} but not in open surgery.\textsuperscript{5} \textsuperscript{19}

In anaesthesia, the ICSAD has potential to objectively assess the development of expertise in technical skills but has not yet been validated for specific anaesthesia procedures and this should be an area of future research. It may also be of use in formative assessment, although the data output is not intuitively interpretable and would perhaps be best compared with previous performance or that of peers. A limitation of the ICSAD is that it can only ever assess process rather than outcome, for instance, it gives no information on whether the procedure was performed well. The data are likely to be more useful if triangulated with a checklist or GRS.

**Contexts of procedural skills assessment**

Procedural skills have historically been both taught and evaluated on patients. In recent years, advances in simulation technology have enabled the evaluation of procedural skills to be taken out of the clinical context into the simulation laboratory. This has created the possibility of standardized assessment of procedural skills.

**Simulation**

Simulation includes the use of manikins, human cadavers, animals, virtual reality, and standardized patients. ‘Part-task’ trainers (PTT) simulate a particular anatomical area or procedure in contrast to full patient simulation with a computer-enhanced mannequin. PTT are generally less expensive than full-patient computerized manikin-based simulators, especially when taking into account the costs of a technician and actors required, but full-patient systems can better simulate the whole clinical environment in order to recreate situations where procedures must be completed quickly to prevent further physiological deterioration.\textsuperscript{88} Hybrid simulation involves more than one type of simulation and can be used to put procedural skills in context. An example is having an actor playing the part of a patient to assess consent and communication skills, with an attached mannequin arm to assess placing an i.v. cannula.\textsuperscript{49} \textsuperscript{50}

The fidelity of a simulation refers to its similarity to an actual task or patient. Although simulation will never be exactly the same as a clinical experience, there are a number of advantages in assessing procedural skills in a simulated rather than a clinical environment. Some procedures such as emergency cricothyrotomy are uncommon enough that most anaesthetists will not perform this procedure during their training. Assessing procedural skills using a simulator prevents potential harm to patients and has been described as an ethical imperative,\textsuperscript{99} although ethical issues remain with the use of animals or human cadavers. In-training assessment has also been challenged as problematic due to power relations between the trainee and the trainer,\textsuperscript{65} potential conflicts of interest between the consultant as a teacher and as an evaluator,\textsuperscript{76} and the trivialization of evaluation and development of a ‘tick box’ mentality.\textsuperscript{6} As there are problems with both standardized simulated evaluations and in-training assessment, a combination of both may offer the best solution.

Predictive validity has been demonstrated for the Human Patient Simulator (METI, Sarasota, FL, USA) for teaching tracheal intubation: 10 h of deliberate practice was found to be as effective as 15 intubations in the operating theatre.\textsuperscript{36} The AirSim (TruCorp, Belfast, Northern Ireland) has content validity as it is anatomically correct as it was designed using a 3D model, using spiral CT scans of the human airway.\textsuperscript{23} The extubated anaesthetized sheep model also has content validity for the ‘can’t intubate, can’t ventilate’ scenario: it has secretions, a mobile larynx, it will bleed if a surgical airway is attempted, it can develop s.c. emphysema, it has a cricothyroid membrane, and it desaturates after extubation producing a sense of responsibility to regain the airway. There are, however, both cost and ethical disadvantages to this model.\textsuperscript{38} Construct validity has been established for some simulations in anaesthesia including the Bill 1 airway simulator (VBM Medizintechnik GmbH, Sulz, Germany) for cricothyrotomy\textsuperscript{82} and a virtual reality flexible bronchoscopy simulator (Immersion Medical, Gaithersburg, MD, USA).\textsuperscript{15} However, the majority of studies on the validity of anaesthesia simulators for procedural skills have generally limited themselves to discussions of face validity, rated by subjective realism by the users.\textsuperscript{89} Face validity depends largely on the simulation’s fidelity. Although there is some evidence that low-fidelity PTT are as effective as high-fidelity simulations in the teaching of anaesthesia procedural skills,\textsuperscript{12} \textsuperscript{56} this has not been investigated for evaluation.

Surgical specialities are considerably ahead of anaesthesia in the development and validation of PTT. One example is the McGill Inanimate System for Training and Evaluation of Laparoscopic Skills (MISTELS), a non-anatomical simulator that has been assessed at multiple institutions.\textsuperscript{27} The MISTELS simulation and metric has
been found to be reliable with excellent internal consistency, inter-rater and test–retest agreement, construct validity, and concurrent validity.

There is good evidence that simulation-based learning with PTT is effective for teaching procedural skills in surgery. Virtual reality PTT, both non-anatomical and anatomical, have been demonstrated to have predictive validity for the teaching of laparoscopic cholecystectomy and reduce error when operating on patients. Virtual reality PTT have also been used for high-stakes evaluation in surgery. Predictive validity has only been demonstrated in anaesthesia for the teaching of tracheal intubation. However, procedural skills in anaesthesia can be as vital as decision-making out of the operating theatre can be in the surgical specialities and it could be argued that PTT for anaesthesia have been neglected compared with full-patient manikin simulators. As an example, it is notable that there is no commercially available airway PTT that is capable of simulating a comprehensive range of pathology that can cause a difficult airway.

The success of outcome-driven simulation-based learning suggests that using simulation as an evaluation before independent practice has the potential to impact positively on patient safety. At the present time, there has been no research that has demonstrated predictive validity for any evaluation using any simulation. A priority for future patient safety research should be to demonstrate that a competent performance on a simulator results in competence in actual patients.

**Multiple station examinations of procedural skills**

The Objective Structured Assessment of Technical Skills (OSATSSs) were developed as an objective assessment of procedural skill outside of the operating theatre and is similar to an OSCE. Candidates perform a series of standardized surgical skills on bench models. At each, time-limited station, candidates are examined by the direct observation of experts and technical skills are assessed using both a generic GRS and a task-specific checklist. OSATSSs have been shown to be a reliable measure of technical skill. Construct validity and concurrent validity have been established for OSATSS including correlating surgical faculty rankings to OSATSS scores for senior surgical trainees. However, organizing and running an OSATSS examination is labour intensive and relatively expensive. An OSATSS examination format has not yet been used in anaesthesia but has the potential to standardize procedural skills assessment if suitable simulations can be validated, as discussed above.

**Conclusions**

Procedural skills in anaesthesia are assessed poorly compared with other domains of learning. This domain of learning has undergone detailed investigation in surgery because of the high importance of procedural skill on surgical patient outcome and anaesthesia has the opportunity to learn from recent advances in the surgical fields.

It has been argued that evaluation drives learning. Current evaluations in anaesthesia tend to focus on broadly defined technical skills but neglect the details of procedural skills (and non-technical skills). Improving the evaluation of procedural skills has the potential to promote excellence in a neglected domain of learning.

Research into the assessment of technical skills in anaesthesia has been conducted with heterogeneous methodologies, which makes comparison difficult. Future studies should include core elements that ensure methodological rigor and facilitate robust comparisons between trials. Such studies should be able to demonstrate assessments: (i) validity—it measures what it purports to measure, (ii) reliability, (iii) feasibility—including cost-effectiveness, and (iv) comprehensiveness (allows for various levels of difficulty). The demonstration of predictive validity for the evaluation of procedural skills should be considered an achievable priority.

This review has presented a diverse array of different methods each having their own particular advantages and disadvantages. A key question is ‘how should a training programme assess the procedural skills of trainees in practice?’ First, several methods can be excluded. Logbooks and procedure lists are best suited to providing information regarding likely opportunities within training programmes and there is little evidence to promote the use of psychometric or aptitude testing in anaesthesia. Cusum analysis has the potential to provide a robust statistical measure of procedural competence but relies on either self-reported performance or repeated direct observations and can require large numbers of performances to demonstrate competency. New technology such as motion analysis may have a role in focusing on manual dexterity during technical tasks but requires further validation before being used to assess procedural skills in anaesthesia.

Currently, the best evidence for a valid, reliable, feasible, and comprehensive assessment tool to assess procedural skill in anaesthesia lies in the use of checklists and GRSs. There is good evidence for the use of a combination of a checklist and GRS in the setting of medical educational research and this combination of tools could be considered the ‘gold standard’ in this setting. When choosing an assessment tool for procedural skills, the key question is what the purpose of the evaluation is. For formative assessment, checklists and GRSs provide in-depth information to promote learning. Most research has focused on the measurement of performance; however, summative evaluation requires not only the measurement of performance but also the setting of standards and the designation of a trainee as competent or not competent. Subjective judgements of competence as a binary outcome are not psychometrically robust enough for high-stakes purposes and more reliable options include a GRS that
includes competence as a behavioural descriptor or calibrating the scores of an assessment tool against an expert group.

There is a need to investigate available assessment tools in real-time in-training evaluations using large numbers of both trainees and raters. Directorate or faculty development is key: trained staff are more reliable in summative assessment and more likely to give feedback in formative assessment. As Kelly and colleagues have stated: 'a macroview is needed on the purposes and systems for assessment, before becoming involved in the details of developing particular instruments, in particular, although a mediocre assessment within a good system can be manipulated to be "good enough", a good instrument within a poor system is likely to perform poorly'.

Patient simulation is a growth area in anaesthesia training programmes; however, simulation for evaluation remains a controversial area. Research is required to identify such systems. There is not enough evidence to recommend that simulation and multistation assessments similar to OSATS may become a key part of the future of procedural skills assessment in anaesthesia if suitable part-task simulators can be validated. However, at the present time, there is not enough evidence to recommend that anaesthetic trainees are evaluated in procedural skills using simulators, especially when these skills can be reliably assessed by direct observation of performance on patients.

**Appendix**

Formulæ for calculating the variables for Cusum analysis:\(^{46}\)

\[
\begin{align*}
    a &= \ln \left[ \frac{1 - \beta}{\alpha} \right] \\
    b &= \ln \left[ \frac{1 - \alpha}{\beta} \right] \\
    P &= \ln \left( \frac{p_1}{p_0} \right) \\
    Q &= \ln \left( \frac{1 - p_1}{1 - p_0} \right)
\end{align*}
\]

where \(h_0\) is the upper boundary limit \(-b/(P+Q)\); \(h_1\), the upper boundary limit \(-a/(P+Q)\); \(s=Q/(P+Q)\); \(\alpha\), the risk of a type I error; \(\beta\), the risk of a type II error; \(p_1\), the acceptable failure rate; and \(p_0\), the unacceptable failure rate.

**References**

15. Crawford SW, Colt HG. Virtual reality and written assessments are of potential value to determine knowledge and skill in flexible bronchoscopy. *Respiration* 2004; 71: 269–75
88 Vazdodaria BS, Gandhi SD, McInoido AK. Comparison of four different emergency airway access equipment sets on a human patient simulator. Anaesthesia 2004; 59: 73–9
97 Xeroulis GJ, Park J, Moulton CA, Reznick RK, LeBlanc V, Dubrowski A. Teaching suturing and knot-tying skills to medical students: a randomized controlled study comparing computer-based video instruction and (concurrent and summary) expert feedback. Surgery 2007; 141: 442–9