Chapter 8
Metric-Based Training to Proficiency: What Is It and How Is It Done?

In countries such as the UK and Ireland, high profile medical error cases had profound implications for the process under which doctors were deemed qualified to practice medicine. It also brought to the fore once again the debate about competency.

Competence Definition and Assessment

A common definition of competence is: "the condition of being capable; having *sufficient* skill and/or knowledge; *the state* of being legally competent or qualified" (Dictionary 1995). Another definition of competence is "the *minimal* level of skill, knowledge, and/or expertise derived through training and experience, required to safely and proficiently perform a task or procedure" (Marshall 1995). Pitts et al. (2006) note that there are debates about the nature or meaning of the word competence. One conceptual standpoint states that a competence is simply a demonstrable ability to do something, using directly observable performance as evidence. Another understands competence as being a: "holistic integration of understandings, and professional judgments, where 'competence' is not necessarily directly observable, rather it is inferred from performance."

One of the problems with the above definitions is that they are not really definitions but mere descriptions. In Chap. 4, we discussed the issue of operational definitions which are a pre-requisite for measurement of performance and in addition these definitions must be refutable. Falsifiability or refutability is the logical possibility that an assertion can be shown false by an observation or a physical experiment. That something is "falsifiable" does not mean it is false; rather, that if it is false, then this can be shown by observation or experiment. The term "*testability*" is related but more specific; it means that an assertion can be falsified through

A.G. Gallagher and G.C. O'Sullivan, *Fundamentals of Surgical Simulation*,
Improving Medical Outcome - Zero Tolerance,
DOI 10.1007/978-0-85729-763-1_8, © Springer-Verlag London Limited 2012

experimentation alone. These descriptions of competence only give clues as to how competence might be assessed but do not specify what “capable,” “sufficient,” or “minimal” might mean in real terms. Falsifiability is a very important concept in *science* and the *philosophy of science*. The concept was most clearly expounded by *Karl Popper*. He concluded from his philosophical analysis of the *scientific method* that a *hypothesis*, *proposition*, or *theory* is “scientific” only if it is falsifiable (Popper 1979). This makes “competency” difficult to assess. Another problem with the understanding of the concept of competence is well demonstrated by Pitts et al. (2006). Competence is not as stated by them the capacity to demonstrate the ability to do something; rather it is the ability to demonstrate doing something to a certain standard.

Medicine has developed a wide array of techniques to assess the “competence” of doctors in training and it is from the results of these assessments that competence is inferred. The majority of these tests assess medical knowledge. However, in the 1970s, there was a move away from just assessing what the medical trainee knows, to what they could do. In 1963, Howard S. Barrows introduced the “standardized patient” into medical education and training (Barrows and Abrahamson 1964). The first standardized patients were, in fact, out-of-work Hollywood actors who were employed by the University of Southern California to play the role of patients. Playing the role of a real patient meant that each student had an opportunity to come face-to-face with the totality of the patient, his stories, physical symptoms, emotional responses to his ailments, attitudes toward the medical profession, stresses with life, work, and family. In essence, the standardized patient brought everything to the clinical situation that a real patient brings. The theory behind the practice was that the student could experience and practice clinical medicine without jeopardizing the health and welfare of a real, sick patient. The term standardized patient became adopted and widely used during the 1980s by medical education researchers who were primarily interested in clinical evaluation of performance.

In the UK, there was also considerable concern about how to assess clinical competence. Clinical competence was usually assessed by two examiners who tested the trainee’s skill on a few patients. Thus, the luck of the draw played a major part in the procedure and variation in the marking standards between examiners was also a problem. Also, frequently, there was confusion about what was being tested, e.g., from being a test of skills in eliciting a history or carrying out a physical examination and a history to a test that was more about the candidates’ factual knowledge than their clinical skills. In response to these problems, Harden and colleagues from the Department of Medical Education, University of Dundee developed the objective structured clinical examination or OSCE (Harden et al. 1975). In the structured clinical examination, the variables and complexity of the examination were more easily controlled. Other advantages that the OSCE had over the more traditional assessment was that it had clearly defined aims, which meant that more of the candidates’ skills could be assessed with a more objective marking strategy which had been agreed with assessors in advance. The object of the OSCE is to assess basic and clinical skills in a reliable format.

It is a flexible test format based on a circuit of patient stations. At each station, trainees interact with a real patient or a standardized patient to demonstrate specific skills. These stations may be short, e.g., 5 min or long, e.g., 15–20 min and there may be as few as eight stations or more than 20. Scoring is done with a task-specific checklist, rating scales, or a combination of a checklist and rating scale. Scoring can be done by the assessors or by the standardized patients. The designing of an OSCE is usually the result of a compromise between the assessment objectives and the logistics constraints, but the content is always linked to the curriculum. If the OSCE scorers are being used for making a pass-fail decision, then it is necessary to set standards and scores. OSCEs are based on tasks that approximate performance in the clinical area of interest and the assumption is that the closer the tasks are to the clinical reality the more valid the assessment. However, there are a number of problems with this approach. The first is that each station is time limited, and so only allows trainees to perform isolated aspects of the real clinical situation. This fragmented approach provides a better opportunity to assess more performance characteristics of the trainee however; this is at the cost of degrading the doctor–patient encounter. The task-specific checklist assessment procedure for the OSCE has also been criticized. It is been proposed that checklists tend to emphasize thoroughness and may become less relevant as the experience of the candidate increases.

Assessment is like good science; once you know the questions to ask, development of the experimental design to answer the question is relatively straightforward. Medical education tends to have the same problem and once it has worked out what it should be assessing it sets about developing a sound assessment strategy. Bryant (1969) has said "examinations are about the least understood and most misused tools of education. They are used mainly to certify that the student has learned an acceptable amount of what he has been taught and to provide a grade representing that attainment. While the announced objectives of the institution may be to develop the knowledge, skills, and attitudes necessary to being a good doctor, the examination seldom measures more than the simple recall of isolated pieces of information. The student grade is usually determined by comparing their performance with the class as a whole; that is, 'grading on the curve' rather than grading according to standards carefully developed by the faculty (p. 209, 1969)." What Bryant is suggesting is that in the assessment of medical skills, the goal should be the assessment of competence rather than just assessment per se.

Competency: Accreditation Council for Graduate Medical Education

The Accreditation Council for Graduate Medical Education (ACGME) is responsible for the *Accreditation* of postgraduate medical training programs within the USA. In response to growing criticism of graduate medical education from a variety of

Table 8.1 ACGME core competencies

Competency	Definitions
1. Patient care	Provision of *timely*, effective, appropriate, and compassionate patient care
2. Medical knowledge	Uses medical knowledge for clinical problem solving and decision making Able to identify life-threatening conditions Able to formulate an appropriate differential diagnosis
3. Interpersonal and communication skills	Able to conduct an effective information exchange with patients, their families, and medical colleagues
4. Professionalism	Arrives on time, ready to work Maintains a proper appearance Inoffensive dress and appropriate cleanliness Appropriate attitudes, respect for patient autonomy, ethical behavior, probity
5. Practice-based learning and improvement	Understands patient care practices and assimilates necessary components for improvement
6. Systems based practice	Capacity to understand, access, and effectively utilize the resources of a given health care system to enable the provision of optimal emergency care

sources (including the medical community itself), the ACGME identified general competencies which all graduates should be able to meet on completion of their training (Beall 1999). The criticisms of graduate medical education center around the fact that many medical trainees were not adequately prepared to practice medicine in the rapidly changing healthcare environment. The core competencies that the ACGME developed are given in Table 8.1.

The ACGME explains in detail what performance characteristics contribute to and constitute specific competencies. It is the responsibility of a training program and the trainees to ensure that competencies are demonstrated. The ACGME reassured training program directors that the development of assessment tools would not be the sole responsibility of the training programs and that when validated assessment tools developed by ACGME or individual programs would be made available to all of them. They also assured programs that many of the assessment tools that were being used will almost certainly be appropriate. The most important factor in the continued use of these assessment systems was that they demonstrated to be valid and reliable measures of competency-based learning objectives. Initially, all training programs were encouraged to assess trainee competencies in all six domains with at least one approach in addition to global/end of rotation clinical ratings. Assessment also included direct observation and concurrent evaluation, 360° evaluation involving non-physician members of the care team, patients and families, and checklist evaluation of improvement projects and cognitive tests.

The long-term goal of the ACGME was to develop a new model of accreditation that was directly linked to the six general competencies. Furthermore, because the competencies were created in conjunction with the American Board of Medical

Specialties, it was hoped that this model of certification could be used in an ongoing basis for accreditation of physicians throughout their careers. The new competencies model was seen as a potential solution to the exponential increase in training requirements in medical education in the USA. Competency-based training offered a more innovative approach rather than the traditional prescription of what was required to be considered medically trained. However, that is not quite what has come out of the ACGME competencies program. Apart from creating general confusion among program directors as to how to achieve or implement a competency program, this new training system has probably created more bureaucracy than it replaced. For example, just some of the assessments that program directors are responsible for include 360° evaluation, chart-stimulated recall oral examination, checklist evaluation of the live or recorded performance, and global rating live or recorded performance, OSCE, Procedure, Operative, or Case logs, patient surveys, portfolios, simulations and models, standardized oral examination and written multiple choice questions (MCQ's).

Lurie et al. (2009) in a systematic review of research on the ACGME, six general competencies found that between 1999 and March 2008, 127 articles were published of which 56 met their specific review inclusion criteria (i.e., validation studies or instrument development). They found that quantitative studies of evaluation failed to develop measures reflecting the six competencies in a reliable and valid way. Overall, they concluded that the research literature provides no evidence that current measurement tools can assess the competencies identified by the ACGME independently of one another. The exception to the challenge of measuring competency was medical knowledge; measures which reliably assess medical knowledge seemed to be valid predictors of important clinical performance characteristics. This finding does not really come as a surprise as the assessment of medical knowledge has been a pillar of medical education almost since its inception. By contrast, the other five competencies reflect in varying degrees personal attributes of trainees rather than knowledge of objectively derived information. Furthermore, the relative value of these attributes is more socially and culturally determined than they are of education and training. Even concepts such as "professionalism" which predated the ACGME general competencies have "continued to defy a clearer operational definition despite several decades of attempts to derive one" (p. 306). To compound these stark conclusions is the fact that one of the specifically recommended assessment strategies proposed by the ACGME (Assessment Toolbox) is OSATS. In Chap. 7 we have explained in some detail why the published evidence on OSATS fails to meet an acceptable level of reliability for use in high stakes assessment. Overall, one of the major problems with the competencies proposed by the ACGME is that they have offered extensive detailed descriptions of what constitutes specific competencies; however, they have offered few if any operational definitions. For example, the *Practice-based Learning and improvement competency states that the trainee* "Understands patient care practices and assimilates necessary components for improvement." How is this competency falsifiable; what is it that the trainee must do, to whom and how frequently before the program director or educational supervisor decides that they do not meet this competency? Without precise operational definitions, it is not possible to reliably and validly assess performance. Simply working from the descriptions of the competencies described by the ACGME it would seem a herculean task to try and develop valid and reliable assessment tools. Lurie et al. (2009) quite sensibly conclude and recommend that the competencies identified by the ACGME should be used to guide and coordinate specific evaluation efforts rather than attempting to develop instruments to measure the competencies directly.

Competency: United Kingdom, Canada, Australasia, Ireland

Training programs in the UK, Canada, Australasia, and Ireland were under the same pressures as in the USA to examine their training and assessment practices for doctors. Rowley and colleagues (Pitts et al. 2006) stated that although the job of a surgeon cannot be neatly defined, it can at least be broken down into a series of outcomes that would lend themselves to assessment. On the matter of professionalism, the GMC detailed what it considers the constituent parts of this attribute in "Good Clinical Practice" (The principles of good clinical practice are outlined in articles 2–5 in the EU Directive 2005/28/EC (Verheugen 2005)) and Tomorrow's Doctors (General Medical Council 1993). However, Rowley (Pitts et al. 2006) from Ninewells Hospital, Dundee, Scotland suggests that the attributes of a surgeon are better captured in the work of the Canadian Medical Association outlined in their CanMED 2000 project (Frank 2005) and these are detailed in Table 8.2.

The CanMeds project suggested that competencies are "…important observable knowledge, skills and attitudes" that they chose as the central concept in planning medical education in Canada. This reflected the ultimate goal of the CanMEDs project which was to develop the abilities of physicians needed to provide the highest quality of care. The process of identifying the core abilities involved translating the available evidence on effective practice into educationally useful elements. The result was a new multifaceted framework of physician competence that comprised a number of competencies. To be useful, these were organized thematically around "meta-competencies" or physician Roles for CanMEDS (outlined in Table 8.2). Traditionally medical education has articulated competence around core medical expertise. In the CanMEDS construct, Medical Expert is the central integrative role but is not the only one. Domains of ability that have long been described or displayed by the effective physician were made more explicit and re-emphasized and articulated as a specific goal of training (Aretz 2003; Epstein and Hundert 2002; Neufeld et al. 1998).

The first step in the process of implementing these aspirations was to devise a curriculum that comprehensively detailed the qualities required and these were traceable back to categories in the CanMEDs 2000 for the nine major disciplines of surgery. One of the major parts of this curriculum was the required assessment methodologies. In the past, the knowledge and judgment of surgical trainees was

Table 8.2 CanMEDS roles and definitions

Roles	Definitions
1. Medical expert	As medical experts, physicians integrate all of the CanMEDS roles, applying medical knowledge, clinical skills, and professional attitudes in their provision of patient-centered care. Medical expert is the central physician role in the CanMEDS framework.
2. Communicator	As communicators, physicians effectively facilitate the doctor–patient relationship and the dynamic exchanges that occur before, during, and after the medical encounter.
3. Collaborator	As collaborators, physicians effectively work within a healthcare team to achieve optimal patient care.
4. Manager	As managers, physicians are integral participants in healthcare organizations, organizing sustainable practices, making decisions about allocating resources, and contributing to the effectiveness of the healthcare system.
5. Health advocate	As health advocates, physicians responsibly use their expertise and influence to advance the health and well-being of individual patients, communities, and populations.
6. Scholar	As scholars, physicians demonstrate a lifelong commitment to reflective learning, as well as the creation, dissemination, application, and translation of medical knowledge.
7. Professional	As professionals, physicians are committed to the health and well-being of individuals and society through ethical practice, profession-led regulation, and high personal standards of behavior.

assessed by summative methods, e.g., MCQs, essays, viva's or orals, and clinical examinations. In many respects, surgeons have always assessed trainees in the workplace because of the apprenticeship tradition. However, some of the problems with this approach through the years have been the lack of objectivity and some surgeons felt that undue influence may have been too important a factor in the assessment of some trainee surgeons. Nevertheless, workplace assessment offers great opportunities if the issues of reliability and validity can be resolved. Assessment tools that were developed specifically to resolve these issues were Direct Observation of Procedural Skills (DOPS) and Norcini et al. (2003) mini-CEX which could be applied in everyday situations in real-time. In the workplace, assessment tools need to be practicable as well as valid and reliable. This means that assessments should be brief and focused on small areas of activity which should limit the effect on a busy working hospital whilst capitalizing on the relevant environment. For example, during a surgical attachment, a young trainee may agree with his trainer that by the end of the attachment, he should be proficient at hernia repair. After a number of months of gradually doing more and more (and after a series of formative assessment sessions), the trainee is ready to be assessed. All the learning objectives are found to have been met, and after a 10 min debrief at the end of an operation, the trainee and the trainer agree that the trainee has demonstrated the key competence. This would be repeated in different attachments with other trainers and gradually a body of evidence from different assessors is accumulated into a growing competence portfolio.

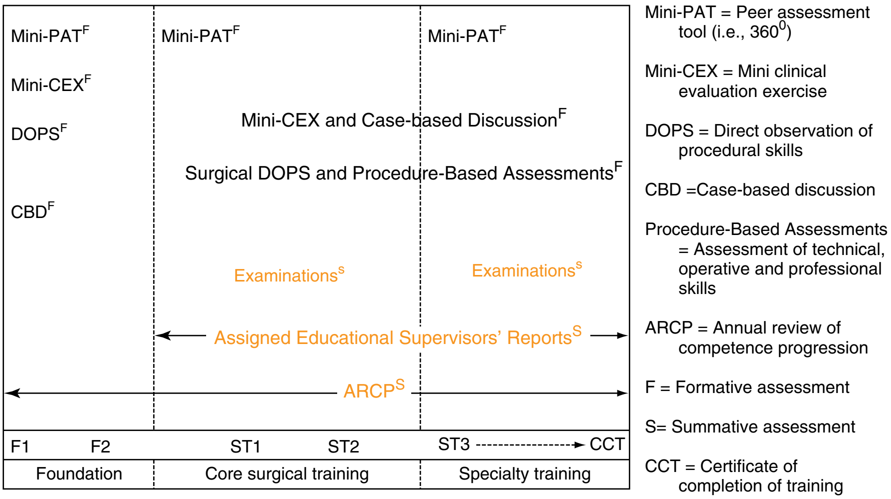
Mini-PAT^F
Mini-CEX^F
DOPS^F
CBD^F
Mini-PAT^F
Mini-PAT^F
Mini-CEX and Case-based Discussion^F
Surgical DOPS and Procedure-Based Assessments^F
Examinations^S
Examinations^S
Assigned Educational Supervisors' Reports^S
ARCP^S
F1
F2
ST1
ST2
ST3 → CCT
Foundation
Core surgical training
Specialty training
Mini-PAT = Peer assessment tool (i.e., 360^0)
Mini-CEX = Mini clinical evaluation exercise
DOPS = Direct observation of procedural skills
CBD =Case-based discussion
Procedure-Based Assessments = Assessment of technical, operative and professional skills
ARCP = Annual review of competence progression
F = Formative assessment
S= Summative assessment
CCT = Certificate of completion of training

Fig. 8.1 Intercollegiate Surgical Curriculum Programme (ISCP) for training and assessment

The type of assessment depends on the stage of training of the individual. These are shown in Fig. 8.1 when the trainee enters into training at Foundation One (or F1) through Core surgical training (ST1 up to STn (can be ST7 or ST8)) and ends on receiving the Intercollegiate Surgical Curriculum Programme (ISCP) Certificate of Completion of Training (CCT). Most of the assessment process is formative, but the annual review of competence progression (ARCP) and assigned educational supervisors reports and exams are summative.

ISCP Assessments Contributing to Competency Assessment

Mini-PAT

The mini-PAT assessment is sometimes described as the 360° assessment or multisource feedback. It is a method of assessing professional competence within a team working environment and providing development feedback to the trainee. It is first undertaken at entry-level (F1) and then every 3 years in specialty training and more frequently if there are concerns. Trainees are expected to understand the range of rules and expertise of team members in order to communicate effectively to achieve high-quality service for patients. Mini-PAT comprises a self-assessment and trainee performance assessment from a range of co-workers (range 8–12) who are chosen by the trainee and will always include the assigned educational supervisor. The assessment will not include administrators or patients. The competencies assessed map across to the Standards of Good Medical Practice and to the core objectives of the intercollegiate surgical curriculum. The assigned educational supervisor signs off on the trainee's mini-PAT assessment and makes comments for the annual review.

Mini-CEX

The mini clinical evaluation exercise (or mini-CEX) is a method of assessing skills essential to the provision of good clinical care and to facilitate feedback. It assesses the trainee's clinical and professional skills on the ward, on ward rounds, in Accident and Emergency and in outpatient clinics. Trainees are assessed on different clinical problems that they encounter in a range of clinical settings. Trainees should choose different assessors for each assessment, but one assessor must be their assigned educational supervisor. Assessors must be registered with ISCP and have expertise in the clinical problem on which the trainee is being assessed. The assessment involves observing the trainee interact with the patient in a clinical encounter. The areas of competence covered include: history taking, physical examination, professionalism, clinical judgment, communication skills, organization, efficiency, and overall clinical care. They normally take between 15 and 20 min with the patient and 5 min afterwards with the assessor. Mini-CEX should be undertaken at least six times per year in specialty training years ST1 and ST2. Their use in specialty training will depend on the specialty and level of training.

DOPS

Direct observation of procedural skills (or DOPS) is used to assess trainee's technical, operative, and professional skills in a range of basic diagnostic and interventional procedures, or part procedures during routine surgical practice. Surgical DOPS are used in some environments and procedures and can take place in wards, outpatient clinics, and the operating theater to facilitate developmental feedback. The original DOPS was developed by the UK Royal College of Physicians. The surgical DOPS can be used routinely every time the trainer supervises a trainee trying out one of the specified procedures, with the aim of making the assessment part of routine surgical practice. The assessment involves an assessor observing the trainee perform a practical procedure and then evaluating performance on a structured checklist that enables developmental feedback to the trainee immediately afterwards. An overall rating on any one assessment can *only* be completed if the entire procedure is observed and judgment will be made at the completion of the rotation as to the overall performance level achieved in each of the assessed surgical procedures. Surgical DOPS should be undertaken at least six times per year in ST1 and ST2.

CBD

Case-Based Discussions (CBD) were designed to assess clinical judgment, decision making, and the application of medical knowledge in relation to patient care in cases for which the trainee has been directly responsible. As such, the method was designed to test higher order thinking and synthesis and allows the assessor to observe how the trainee elicits, prioritizes, and applies knowledge. The function of the exercise is not focused on the trainee's ability to make a diagnosis; rather, it is more like a structured in-depth discussion between the trainee and their assigned educational supervisor about how the managed a clinical case. Challenging cases are preferred as this allows a trainee to explain the complexity involved and the reasoning behind the choices they made in the care of that patient. It also facilitates discussions on the ethical and legal parameters of clinical practice. Real patient records form the basis for dialogue, systematic assessment and structured feedback. This also allows the assessor to evaluate the quality of the trainee's recordkeeping and presentation. Assessments usually take about 15–20 min, followed by 5 min of feedback from the assessor.

Procedure-Based Assessments

Procedure-Based Assessments (PBAs) are used to assess a trainee's technical, operative, and professional skills in a range of specialty procedures or part of procedures during routine surgical practice. These provide a framework to assess practice and facilitate feedback in order to direct learning. The assessment method uses two principal components. The first is PBA form for the assessment of a series of competencies within six domains. These are content, preoperative planning, preoperative preparation, exposure and closure, intraoperative technique and postoperative management. Each one of the competencies is assessed with a number of performance characteristics, e.g., for exposure and closure these include:

E1. Demonstrate knowledge of optimum skin incision
E2. Achieved an adequate exposure through purposeful dissection in the correct tissue planes and identifies all structures correctly
E3. Completes a sound wound repair
E4. Protects the wound dressings, splints, and drains
E5. See specific PBAs

Each one of these performance characteristics is scored as, N = Not Observed or Not Appropriate; U = Unsatisfactory; and S = Satisfactory. The procedure chosen to be assessed should be representative of those that the trainee would normally be expected to be able to carry out at their level and will be one of a list of index procedures relevant to the specialty. Usually the assessor will be the trainee's assigned educational supervisor but other surgical consultants should also complete the assessments. Trainees should complete assessments on as many procedures as possible with a range of different assessors. During the assessment, the assessor can provide verbal prompts and if required intervene if patient safety is at risk.

PBAs have been adopted as the principal method of assessing surgical skills, the combined competencies specific to the procedures with generic competencies such as safe handling of instruments. They cover the entire procedure, including preoperative and postoperative planning. PBA forms have been developed for all the links procedures in all surgical specialties. The forms were designed to be quick and easy to use as assessments should be as frequent as possible when performing index procedures as a primary aid to learning. PBAs focus on index procedures in each specialty and should be used every time the index procedure is performed.

ARCP

The Annual Review of Competence Progression (ARCP) is a formal review of how well a trainee is progressing in relation to their learning agreement for their training program including their ability to go to the next level. The ARCP is underpinned by appraisal, assessment, and annual planning. The panel bases their decision on the evidence submitted by the trainee and record or the competencies attained and their progression through the training program. The ARCP panel of assessors may include the Training Program Director, other members of the relevant Specialty Training Committee, a College representative, a Deanery representative, an academic representative, an "external" representative, or a lay representative. The ARCP panel reviews a trainee's progress based on the evidence submitted and provides the trainee with an outcome. The panel is explicit about what trainees are required to submit for their review but this will include:

- Structured reports from their Educational Supervisor
- College Assessment Forms (via the ISCP)
- Clinical Logbook (via the ISCP)
- Portfolio
- (Updated Registration Form (Form R))

The outcome of the ARCP will determine the rate at which trainees progress through the training program. Possible outcomes include, incomplete evidence presented (more training time required), released from training without specified competences, inadequate progress, development of specific competencies required, and satisfactory progress, and if trainees consistently underperform or fail to supply sufficient evidence to ARCP, they may be asked to relinquish their National Training Number. The ARCP also provides a mechanism for determining certificate of completion of training (CCT) dates for trainees.

IRCP Assessments Assessed

Overall the Intercollegiate Surgical Curriculum Programme (ISCP) has done an excellent job in constructing a systematic, evidence-based, and targeted training program. Like the ACGME competencies program, they have set out the training for the performance characteristics that a well-trained doctor should possess. They have highlighted "softer" but important aspects of being a good doctor and made it clear that they are as much part of what is being assessed as medical skill. The ISCP has been much more rigorous in what they will accept as assessment of competencies in comparison to the ACGME. It is very impressive that the ISCP has PBSs already developed for every index surgical procedure. ACGME appears to be less advanced in its assessment efforts.

However, the ISCP competency assessment is not without problems. Although performance had been designed to be user-friendly, the entire process seems exceptionally bureaucratic. Perhaps this is the price that has to be paid for objectivity, transparency, and fairness in medical training at the start of the twenty-first century! The emphasis throughout training on formative feedback is good educational practice and optimizes learning opportunities for the trainee; however, there is still considerable room for subjectivity to creep into the system. This is particularly the case for the DOPS because of the Likert-scale, which we have discussed in Chap. 7. It is very difficult to achieve a high inter-rater reliability when using a Likert-scale and good inter-rater reliability levels (i.e., >0.8) are a fundamental component of a valid assessment. The PBAs are certainly one of the most impressive aspects of the ISCP assessment process, particularly as it is procedure-specific for index operations. However, the assessment metrics could certainly be made much more explicit and operational definitions of performance characteristics could be made much tighter. Definitions of performance characteristics such as optimum (without definition), adequate, sound, and purposeful leave too much room for individual interpretation and almost certainly will impact on their inter-rater reliability. However, the Intercollegiate Surgical Curriculum Programme (ICSP) has wisely not gone down the road of Likert-scale type assessments for PBAs and has instead opted for the more robust assessment process where the assessors are simply asked to assess whether performance was unsatisfactory or satisfactory. These problems are not insurmountable and will be addressed in Chap. 12.

Somewhat more worrying about the ISCP assessment systems are their definitions of the meaning of valid and reliable;

- *Valid* – To ensure face validity, the workplace-based assessments comprise direct observations of workplace tasks. The complexity of the tasks increases in line with progression through the training program. To ensure content validity, all the assessment instruments have been blueprinted against all the Good Medical Practice/CanMEDS domains.
- *Reliable* – In order to increase reliability, there will be multiple measures of outcomes. ISCP assessments make use of several observers' judgments, multiple assessment methods (triangulation), and take place frequently.

These could be put forward as one set of definitions but as discussed in Chap. 7 these are not the conventional definitions of "valid" and "reliable" in the context of assessments, particularly when used for high stakes decisions. This issue will almost certainly come under close scrutiny if a trainee who has been failed by this system chooses to challenge it legally. Another problem with the assessment process particularly in the PBAs is what constitutes "satisfactory." We assume that some type of construct validation has been conducted on the individual index PBA assessment procedures to guide this decision making. However, these studies have not been reported in the literature. Another question which needs to be addressed by the ISCP assessment system directly relates to the issue of competency; how many times must a procedure be conducted and assessed as satisfactory for the trainee to be defined as competent? Furthermore, like the ACGME, there is extensive discussion about competence and competencies but at no point does the ISCP operationally

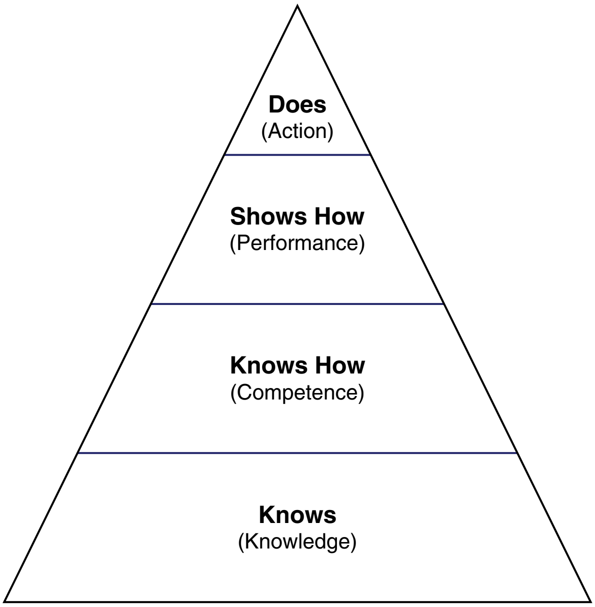

Fig. 8.2 Miller's triangle framework for clinical assessment

defined what they actually mean by competence. They give extensive descriptions of what they consider competent or what competencies are, but they do not offer a definition which is falsifiable.

Competency: Millar's Triangle Model (1990)

George E. Miller (1990) when asked to address the issue of assessment of clinical skills, competence, and performance concluded that no single assessment method could provide all the data required for a judgment of something so complex as the delivery of professional services by a successful physician. He used a triangle/pyramid model to illustrate how he construed the coalescence of performance characteristics that made a successful physician (shown in Fig. 8.2). At the base of this process is knowledge; that is, the trainee physician knows what is required in order to carry out their professional functions effectively. The trainee must also know how to use the knowledge that they have accumulated. They must develop among other things, the skill of acquiring information from a variety of human and laboratory sources. Having acquired this information, they must then be able to analyze and interpret this information so as to formulate a diagnosis and then a treatment plan. It is having sufficient knowledge, judgment, and skill that define competence (according to Webster's dictionary). Traditionally, these qualities and attributes have been assessed with medical exams. However, Miller (1990) points out that traditional academic exams failed to accurately represent how the trainee might deal with an actual patient in vivo rather than in academic examination exercise. He suggests that it is not enough for a trainee to know the way that something is done to be considered competent but they must also show how it is done. Of course one of the challenges that this question poses to the academic clinical community is how to conduct a reliable and valid assessment of that performance. Although considerable advances have been made in the assessment of clinical performance, e.g., standardized patients and OSCE's, the question remains whether what is done in the artificial examination setting is an accurate reflection and good predictor of what a successful medical graduate does when functioning independently in clinical practice. Although we have highlighted that some of the problems associated with the construction of Procedure-Based Assessments, we believe that they are a natural evolution of an optimal assessment process. They have considerably more strengths than weaknesses and we believe will prove a reliable predictor of mature clinical performance.

The problem with Miller's formulation of competency it is that is just like the other approaches we have outlined already; it simply restates the problem and reminds us how difficult it is to measure. He does not offer a definition of competence that is refutable. Furthermore, Miller (1990) appears to assume that as knowledge testing plays such a crucial role in medical education and training progression, success in overcoming that hurdle is by default, an indication of competency. Miller explicitly presents this assumption in his original paper where he aligns "KNOWS HOW" with Competence (p. S63). In reality, even in 1990, this almost certainly was not the case. Medicine to a large extent is a learned skill, and the assumption probably was that these skills were acquired at the same rate as the knowledge of how and when to practice them. High-profile medical error cases in medicine around the world have cast considerable doubt on that assumption to the point where these skills are now explicitly assessed, hence the discussion of competency. Traditionally, medical knowledge has been very well assessed, unlike medical skills. Compounding this problem is the fact that medical education practitioners now know that the same scientific and philosophical (and effort) underpinning of medical knowledge assessment and validation must also be applied to learning, assessment, and validation of procedural skills. It is no longer acceptable to assume that by the time physicians have completed their training, they will have sufficient skills to practice medicine safely. This still leaves the problem of what is sufficient?

Competency as Part of a Skill Acquisition Process

A more comprehensive account of the skill acquisition process has been proposed by Dreyfus and Dreyfus (Dreyfus et al. 1986, 2000). Both brothers were academics at Berkeley; Hubert was a professor of philosophy in the graduate School and his brother Stuart was an applied mathematician. They proposed their theory in direct opposition to much of the thinking at the time about the development and applications of computers. Dreyfus and Dreyfus analyzed the difference between

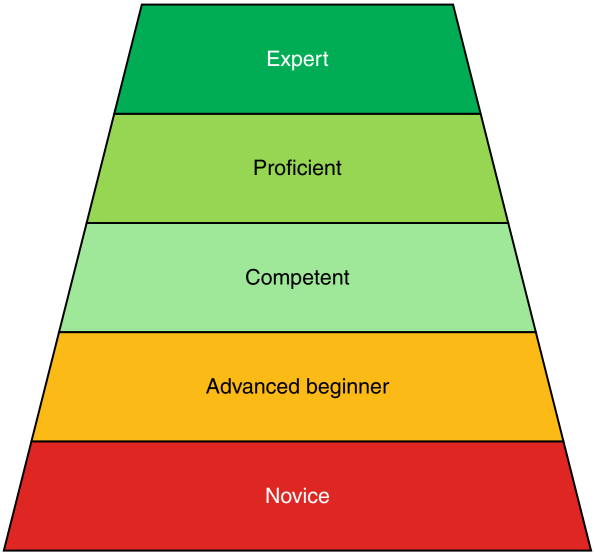

Fig. 8.3 The Dreyfus and Dreyfus (1986) model of skill development which surgery has "embraced"

human expertise and the computer programs that claimed to capture it. They proposed that much of the novelty and intuition that human beings brought to the problem-solving process could not be duplicated or replicated by computers and in particular, they argued against the concept of "computers that can think," or expert machines. In the 1980s, digital computers were basically highly complicated structures of simple switches which were either on or off. The theory on which such machines were based preceded their actual development. Philosophers like Descartes, Pascal, and Leibniz and mathematicians like Boole and Babbage sensed the potential power of combining many simple elements in rule-like ways. By the 1950s, when digital computers were just beginning to be built, logicians such as Alan Turing were already accustomed to thinking of computers as devices for manipulating symbols according to exact rules. The symbols themselves did not mean anything. Computers are general symbol manipulators and so they can simulate any process which can be described exactly.

During the 1950s when digital computers were first constructed, they were first used for scientific calculation. However, by the end of 1950s, researchers like Alan Newell and Herbert Simon began to take seriously the idea that computers were general symbol manipulators. They saw that one could use symbols to represent elementary facts about the world, then use rules to represent relationships between them and then use such rules or programs to deduce how those facts affect each other and what happens when the facts change. In this way, computers seemed to be able to simulate logical thinking. To help inform the discussion about the differences between how machines solve problems and how human beings solve problems, Dreyfus and Dreyfus (Dreyfus et al. 1986) proposed a five-stage model of skill acquisition (which is shown in Fig. 8.3). They were particularly interested in how experts solve problems, the final stage of their model.

Table 8.3 Characteristics of each stage of the Dreyfus skill development model

Stage	Performance characteristics
Expert	• Source of knowledge and information for others • Continually looks for better methods • Work primarily from intuition • Being forced to follow rules degrades performance
Proficient	• Seeks to understand larger context • Frustrated by oversimplification • Can self-correct performance • Can learn from experience of others
Competent	• Can troubleshoot problems on his/her own • Seeks out expert user advice • Develops conceptual models
Advanced beginner	• Starts trying tasks on his/her own • Has difficulty troubleshooting • Begins to formulate principles, but without holistic understanding
Novice	• Has little or no previous experience • Is vulnerable to confusion • Does not know how to respond to mistakes • Needs rules to function

In the model proposed by Dreyfus and Dreyfus (1986), the development of expertise goes through a number of developmental processes from novice through advanced beginner to competence and on to proficiency and then expert. The performance characteristics of each stage of development are outlined in Table 8.3.

There are a number of interesting aspects of the Dreyfus and Dreyfus model. It differs from Miller's model in that they concentrate on what the individual can and cannot do at each stage of skill development. There is also a clear performance hierarchy: from the novice with little or no experience who does not know the rules or how to respond to mistakes, through to the individual who is competent, who has some conceptual models of performance and can troubleshoot some problems on their own but has the insight to seek out expert advice through to the expert who is the source of knowledge and information for others and who continually looks for new and better ways to perform. Another interesting aspect of the Dreyfus and Dreyfus skill acquisition model is that they have subdivided the early parts of skill acquisition into novice and advanced beginner. During the novice stage and the acquisition of new skills, the novice learns to recognize various objective facts and features relevant to the skill and acquires rules for determining actions based upon those facts and features.

At the advanced beginner stage, performance improves to a marginally acceptable level only after the novice has considerable experience in coping with real situations. While that encourages the learner to consider more context-free facts and to use more sophisticated rules, it also teaches them more important lessons involving an enlarged conception of the world, their skill, and the boundaries of their skill capabilities. They start to recognize similar patterns in the presentation of problems and find that the skills (and experience) they have already acquired might help them solve these problems or indeed to at least recognize that they are not equipped to solve the problems. They begin to notice the subtle aspects of their own performance that lead to different outcomes. For example, when driving a needle through tissue, they notice that the angle of entry and the angle path of curvature of the hand that drives the needle through the tissue determine whether they scrape or tear tissues when suturing.

With more experience, the number of recognizable context-free and situational elements present in a real world circumstance eventually becomes overwhelming. A sense of what is important is missing. In general, a competent performer with a goal in mind sees a situation as a set of facts. The importance of the facts depends on the presence of other facts, i.e., context. They have learned that if a situation has a particular constellation of those elements, certain conclusions should be drawn, a decision made, or expectation investigated. They are no longer simply following a set of rules, but begin to perform with a goal in mind. For example, if they are performing a surgical procedure that they have been taught to carry out in a specific sequence or series of steps, they may alter the order of these steps because they believe the new way of performing is more efficient or makes a later part of the procedure easier to perform.

The proficient performer starts to move beyond the position of simply following rules and making conscious choices about goals. A degree of automation becomes apparent in their performance. Automation is the performance of a skill without conscious control (discussed in more detail in Chap. 9) and is usually indicative of a high level of skill acquisition. Although the proficient performer intuitively organizes and understands the task at hand, they still find themselves thinking about what to do. They perform the task in a sequence that they find comfortable, but they readily integrate new and more efficient ways of task performance based on salient aspects of recent performance, i.e., performance feedback (see Chap. 4). They are nearing the top of their learning curve, and in general, performance is tweaked rather than significantly altered.

The expert generally knows what to do based on a mature and practised understanding. Their matured performance which has been honed by experience has by now been largely automated. In Chap. 4 we described how the expert performer needs less attentional resources to perform routine aspects of routine tasks. They appear to perceive and understand the gross and subtle aspects of a case beyond the ability of their less experienced colleagues. Their ability to generate the correct diagnosis with evidence-based reasoning seems almost effortless as does their formulation of alternative treatment plans. These are important aspects of what it is to be an expert. As a general rule, the expanded faculties of being an expert may be considered rather routine during procedures that go routinely. However, when things go wrong during a procedure, the expert has the extra cognitive resources (i.e., attentional), the experience and the skills repertoire to deal with these situations. Dreyfus and Dreyfus (1986) point out that someone at a particular stage of skill acquisition can always imitate the performance characteristics of someone at a higher stage of development when things are going well; however, their true performance level becomes evident when things do not go well. The model of skill and development that Dreyfus and Dreyfus present is a learning model in that skill acquisition passes through distinct stages but the boundaries between these stages are not explicit. Furthermore, learning to perform any task stems from the novice stage of rule governed behavior that then advances to become more automated with experience. The rate of progression will be determined by the talent of a learner, how similar the new tasks are to the performance characteristics and the skills required for previous tasks and also the skill of the teacher.

The Dreyfus and Dreyfus model of skill development has a number of attractive features. The model is intuitively attractive because it is simple and skill acquisition as proposed by them is in a logical uncomplicated sequence. Unfortunately, learning is not that simple as more than a century of quantitative research in psychology and cognitive science has shown. It should also be remembered that Dreyfus and Dreyfus proposed their model of skill acquisition in direct opposition to the proposals of many of their colleagues during the 1970s and 1980s who were suggesting that computers would become intelligent performers of sophisticated human activities. Hubert Dreyfus (1979) argued (and was derided for many years) that human intelligence and expertise depended primarily on unconscious instincts rather than conscious symbolic manipulation and argued that these unconscious skills would never be captured in formal rules. Cognitive science knows considerably more about cognition and cognitive processes at the start of the twenty-first century than they did during the 1980s when the brothers were writing. Instinctive human performance as understood by Dreyfus and Dreyfus is probably more readily recognized as automated performance by cognitive scientists, which is somewhat less mystical than Dreyfus and Dreyfus might have conceived. The other problem with the Dreyfus and Dreyfus model is that it was not developed on the basis of experimental studies (as understood by most experimental psychologists) and so it is non-empirical. In fact, most of their formulation seems to have been based on their experience with nurses at different levels of expertise and chess players.

Proficiency: Beyond Competency

Dreyfus and Dreyfus (Dreyfus et al. 1986) propose something that is quite different from what we have discussed already. Previously we have considered competence as being either present or absent (as proposed by the different medical training bodies around the world). We have also construed it as or different levels of competence (Miller 1990). What Dreyfus and Dreyfus have proposed is that competence represents performance characteristics that are an interim level of skills development between the novice and the expert. Furthermore, the performance characteristics that are attributed to the competent performer on this scale are really not that skilled. They present the performance characteristics of an individual who is really just starting to develop just “enough” skills. While this definition conforms to the dictionary definition it us uncertain that this is the perception of medical competence held by the general public i.e., just enough. A more promising set of performance characteristics is associated with what Dreyfus and Dreyfus call proficient. At this level, the person is starting to act autonomously but at the same time being cognizant of ways to improve their performance. The dictionary definitions of proficiency are:

(a) The quality of having great facility and competence; skillfulness in the command of fundamentals deriving from practice and familiarity
(b) The ability to apply knowledge to situations likely to be encountered and to deal with them without extensive recourse to technical research and assistance

The other attractive feature about the concept of proficiency is that it is not lumbered with the same historical baggage as the concept of competency. The extensive discussion of the concept of competency has resulted in nothing more than numerous elaborate descriptions that have not resulted in closer moves to operational disprovable definitions. Another attractive feature of proficiency is that if one is proficient, one is by default competent as the model proposed by Dreyfus and Dreyfus (1986) holds that skills are developed in a progressive sequence. Although the definitions offered for proficient performance are no better operationalized than those for competence, it is easier to reach agreement on who is demonstrating proficient skills than it is to reach agreement on who is demonstrating competent skills. Even critics of the competence model of skills would agree that the vast majority of senior doctors practicing medicine are at least competent, probably proficient, and some are expert at what they do. This provides a very robust foundation on which to establish a benchmark against which performance can be judged. It means that someone who is considered to be proficient in the practice of their skills is at least competent and at best expert. A good starting point for an operational definition of proficient is “that it is what proficient individuals do.” This definition may not be as elegant as might have been hoped for, but it is very difficult to argue against it. The next task is to measure what it is that individuals who are proficient do. As it turns out, this task is much easier than it might seem.

Proficiency Measured

In Chap. 7 we discussed the different types of validation efforts that were required for the validation of a simulation and the simulation metrics. We also said that one of the most important types of validation that could be undertaken was construct validation. In Chap. 5 we described how metrics were developed from the initial task analysis of the procedure to be learned through to the operational definition of performance characteristics that are associated with performing the task well or badly. If these are indeed valid performance parameters that indicate where on the learning curve someone (novice, trainee or consultant/attending) is performing, we should be able to detect qualitative and quantitative differences between these groups. These performance characteristics or metrics determine how we measure performance, whether it is in the operating room or on a simulator. It may be a single metric unit that distinguishes performance or it may be a conglomeration of metric parameters. For example, Gallagher et al. (2001) found that all of the MIST VR metric measures (time, error, economy of instrument movement (left and right instrument) and economy of diathermy) distinguished between experts and novice performance. This was confirmed by Gallagher and Satava (2002) who assessed the learning curves of experts and trainees. However, Gallagher and Satava also found that the test retest reliability of economy of instrument movement metrics did not reach a satisfactory level of reliability to be used with confidence. Despite this, they still had three robust parameters that reliably measured and significantly differentiated between the performance of experts and novices.

The next step in the scientific validation of these metrics was to establish whether these metrics predicted intraoperative performance. It should be remembered that MIST VR had been widely dismissed by many in the surgical community in the late 1990s as an interesting video game using laparoscopic surgical instruments that looked nothing like performing surgery on a patient. However, the psychomotor performance characteristics and metric measurements of performance had been derived from a task analysis on laparoscopic cholecystectomy by a surgeon, a behavioral scientist, and software engineer. To the untrained eye they may have looked nothing like surgical performance, but on closer scrutiny, the MIST VR tasks were well suited to the job. The starting position for the Yale University team that completed the first VR to OR clinical trial (Seymour et al. 2002) was a virtual reality simulator with well-validated performance metrics. MIST VR performance metrics that were used in this trial were errors and economy of diathermy. Time was excluded as a training metric because the researchers were more interested in training safe performance rather than fast performance. Economy of instrument movement (e.g., how efficiently the instrument was moved from point A to B in real terms) was excluded because of their measurement reliability issues.

There was an extensive and extended discussion within this group about how long or how many trials a trainee should be trained on MIST VR. The researchers came to the same conclusion at the end of each discussion, i.e., all that these training strategies have achieved historically was considerable variability in levels of skills. It was eventually agreed the trainees would train until they reached a benchmark; however, a similar discussion ensued about how the benchmark should be established. The parsimonious solution that was eventually achieved was that the benchmark would be established on the basis of the performance of members of the surgical team who were discussing the problem. After all, the surgeon members of the team were very experienced laparoscopic surgeons, all worked in the same department, all worked with the same surgical trainees, and that all of them recognized that they had a reasonably homogeneous skill set. From previous research, it had been shown that for experienced laparoscopic surgeons, their learning curve on the MIST VR simulator flattened out at about three trials. All of the attending surgeons participating in the study completed five trials on the manipulate-diathermy task on MIST VR on a modified difficult setting. The performance criteria or benchmarks that trainees were to be trained to was established on the basis of the mean score of the attending surgeons on trials four and five for errors and economy of diathermy (for both hands).

Proficiency Benchmarked

It was assumed by the Yale University team that the MIST VR or performance metrics of “errors” and “economy of diathermy” captured important topographical features of the performance characteristics of experienced laparoscopic surgeons. It was the team hypothesis that training a group of trainee surgeons to the benchmark represented by these metrics would impact on skills levels to the extent that there would be transference to intraoperative performance. Although this type of study had not been conducted before in medicine, there was ample evidence from other high skills industries that training in a simulated environment improved performance on a real world task. There was nothing magical or unusual about the metrics that were used to benchmark the experienced surgeon’s performance. These were the metrics that had been demonstrated to be the most reliable and made the most sense, i.e., the goal of the trial was training surgeons to perform the dissection portion of a laparoscopic cholecystectomy using the electrocautery instrument. It should also be noted that the metrics used are like “time” measures, i.e., surrogate measures of skill. However, the difference between the error and economy of diathermy metrics and time is that they more accurately reflected what the trainee was doing on a second-by-second basis and therefore was a good candidate for performance feedback. The goal of training was to help trainees reach a performance criterion level which meant minimizing performance errors and maximizing efficient use of electrocautery. Information on performance errors and inefficient or erroneous use of electrocautery was given to trainees immediately after being enacted. This was achieved by the simulator with an auditory stimulus for electrocautery errors and the virtual tasks turned red to indicate an error had been enacted. As discussed in Chap. 4, augmented feedback of results such as those described here facilitates learning. In simple terms, it tells the trainee that they have just done something wrong as soon as they have done it, which allows them the opportunity to modify their behavior and not make the same mistake in the future. In contrast, a time metric would simply inform them at the end of the task that they had taken too long. This type of information is too ambiguous for the optimal facilitation of learning. If the time metric could be granularized to inform the trainee as to which parts of their performance were taking too long, this would be much better feedback. However, it would still only tell them that they were taking too long and would not give them feedback on the quality of their performance, whereas, feedback on errors and economy of diathermy does.

The mean performance level on MIST VR of the attending surgeons involved in the trial was used as the performance criterion and benchmark because it seemed the most reasonable measure. This was the first time that a performance criterion level was used as a guide for training success in a surgery clinical trial. Possible alternatives might have been using the performance of one surgeon to benchmark performance, using confidence intervals or the more traditional amount of time training or number of trials in training. The traditional approach to training was rejected fairly quickly because of the variability in skills levels. Ironically, the second clinical trial to demonstrate that virtual reality training improves intraoperative performance used precisely this approach, i.e., they trained the virtual reality subjects for ten trials rather than to proficiency (Grantcharov et al. 2004). The results show that the virtual reality–training group performed significantly better than the standard training group; however, this was more by accident than design as this approach to training is inefficient. Training to a benchmark confidence interval was also rejected as the researchers were not sure what the intervals might be based on, i.e., one standard deviation, 1.96 standard deviations, one inter-quartile range, etc. The mean level of the participating attending surgeons' performance was used because it meant that all of them had contributed to the performance criterion definition. It also meant that extreme performances (had they existed) would have been mitigated by better performances. The team were also keen to use the mean performance because they were aware of research that was ongoing in Sweden in the early 2000s which was generating results that the Swedish researchers found difficult to explain. Ahlberg et al. (2007) were investigating the learning curve of trainee surgeons performing Nissen fundoplication. They were particularly interested in whether the trainee surgeons' initial objectively assessed skill levels would be good predictors of the steepness of their learning curves and intraoperative performances. However, what they did find was that the objectively assessed measures of the senior surgeon's skills were the best predictor of their trainee's intraoperative surgical performance. Indeed, it was better than objective assessment of surgical skills on the simulator. The implication of these findings was that the trainees' skills level regress to that of their supervising surgeon (in both directions!).

Choosing the mean performance level in setting the benchmark performance criterion avoids asking difficult questions about surgeons who were not performing as well as some of their colleagues while at the same time establishing a robust skills level that is representative of a given group of surgeons as a whole. If trainees were performing to a benchmark performance criterion level, that meant that their performance was equal to or better than 50% of the performances on which the benchmark was established. Even the most ardent critic of this approach to training would have to admit that this is a much more rigorous approach to training than currently exists. However, there are a number of implications for setting a performance criterion level and how it is established (Chap. 12). The Yale team was also aware that choosing the mean performance of the attending surgeons was probably a conservative approach, but at that time, proficiency-based progression training was an unproven methodology.

Trainees on a proficiency-based progression training schedule continue training on the simulator until they reach the performance criterion level on both metrics, with both hands on two consecutive trials (for the Yale VR to OR trial (Seymour et al. 2002)). The reasons for these specifications were that laparoscopic cholecystectomy is a bimanual task and so it made sense that trainees should be adept at using instruments in both hands. VR allowed training and assessment of bimanual psychomotor performance for the laparoscopic cholecystectomy. Trainees were also required to reach the performance criterion level on both metrics because these were the metrics that best characterized the performance of the attending surgeons. They were required to reach these performance levels on two consecutive trials, because it was argued that they might reach these benchmarks on one trial by accident/coincidence, but it was unlikely that this would be the case for reaching the performance criterion levels on both metrics on both hands on two consecutive trials. Like the issue of mean performance as a benchmark proficiency level, we will return to the issue of proficiency definition in Chap. 12.

One of the advantages of using a virtual reality simulation is that machine-scored performance metrics has been demonstrated to be reliable and valid and takes a lot of the work out of establishing a proficiency level. There is a considerable effort required to develop and then validate the performance metrics but once these have been published, the surgical community can be reasonably confident in their use. However, the problem for surgery is that most of the virtual reality simulators that currently exist are for minimally invasive or image-guided procedures. This approach to surgery continues to represent a minority of surgical practice. The absence of a virtual reality simulator should not impede a proficiency-based progression training program, as demonstrated by the work of Van Sickle et al. (2008). In this clinical trial for training senior residents to perform Nissen fundoplication, no specific virtual reality simulator existed. Instead, the researchers developed novel simulations that captured essential components of the suturing and knot tying required for successful operating. They established performance criterion levels based on experienced operators' performances on these tasks, and then trained surgical residents until they reached these performance levels. The results showed that surgical residents trained to the performance criterion levels performed Nissen fundoplication more efficiently and with significantly fewer objectively assessed intraoperative errors. An important point to note about this study is contained in the discussion section of the paper. They pointed out how time consuming it was to train subjects on a non-virtual reality–based simulation program. It required one of the researchers to observe and in some cases physically score the performance of the trainee while they were training or immediately afterward. However, these are simply implementation obstacles which can be overcome with a determined approach and with innovative solutions. Another important point to note about the Van Sickle et al. (2008) clinical trial is that the researchers went through the same iterative process of metric development, operational definition, construct validation and proficiency definition, proficiency-based progression training, and blinded objective assessment of intraoperative performance to a high level of inter-rater reliability for the outcome assessment. The main point is that proficiency-based progression training quality assures the skills level at the end of the training process, i.e., the graduating trainee is performing as well as or better than 50% of the individuals on whose performance the proficiency levels are established.

Why Proficiency-Based Progression?

Some educationalists may argue that the process that we have just outlined could just as easily be called competency-based progression. However, we disagree. Competency is mired in descriptive detail that is going to make operational definitions difficult to extricate from the baggage. The main problem about competency definition is deciding where the performance criterion line should be drawn. Proficiency does not carry the same baggage. Furthermore, the vast majority of operating surgeons currently in practice operate daily in simple and complex surgical procedures. On the whole, their patients are well looked after, they get good surgical care and safe operative performance. It would be difficult to argue that this is an unreasonable target to set for trainees. The advantage of a proficiency-based approach to training is that we can quantify performance, and in so doing, we set trainees a target that they can reach in their own time-scale. Furthermore, this benchmark is based on something meaningful from the real world, i.e., experienced operating surgeons. For the talented and gifted trainee surgeons, they will reach this target quickly; for the less talented or gifted surgeon, they will take longer to reach the same target but when they do, their skills will be at the same level (at least) as their more talented colleagues. The important point is that they reach this performance criterion level within a reasonable time frame. Will the surgeon who reaches the performance criterion faster become a better surgeon? This is certainly a good research question but current subjective evidence would tend to suggest not. We know from the Yale VR to OR clinical trial team that the resident in their study, who took the longest to reach the performance criterion level performed the best intraoperatively. Furthermore, it takes more than good technical skills to make a complete surgeon.

The Meaning of Proficiency-Based Progression

The apprenticeship model of surgical training has always been credited to the program that Halsted developed at Johns Hopkins in Baltimore, USA. In Halsted's training program (which is not dissimilar to the training program that currently exists in surgery), the trainee was given increasing responsibility for the treatment and care of the patient as their training progressed. Training and progressing were at the behest of the supervising surgeon which of course could be subject to their individual whims. Proficiency-based progression as a training paradigm alters that relationship. Training progression is now determined on a trainee's objectively assessed performance benchmarked against the performance of experienced operators. This means that progression in training is based on objective, verifiable criteria, thus making the process more transparent and fair. Proficiency-based progression training also has implications for the patient. Under the Halstedian training paradigm, the operating room was used as a basic skills training environment where the trainee honed their skills during their training years. In a proficiency-based progression training program, the trainee is not allowed to operate on a patient until they have quantitatively demonstrated that they are performing at the benchmark surgical skills established by their training program. This means that the operating room is no longer a basic skills training environment but more like a finishing school where surgical technique is mastered under the apprenticeship of a senior surgeon.

Proficiency and competency are often used interchangeably; however, they are not the same. In this chapter, we have discussed the differences between proficiency and competency. Proficiency has been operationally and quantitatively defined while competency has only been described, and consequently, there is little agreement among the global medical establishment about the operational definition of competency. Furthermore, precedent has already been established with regard to the quantification and definition of proficiency (Ahlberg et al. 2007; Seymour et al. 2002; Van Sickle et al. 2008). Thus it is prudent to proceed by using a proficiency benchmark as an indicator of skills rather than competency.

Proficiency-based training as a new approach to the acquisition of procedural-based medical skills took a major step forward in April 2004. As part of the roll-out of a new device for carotid artery stenting (CAS), the Food and Drug Administration (FDA) mandated, as part of the device approval package, metric-based training to proficiency on a VR simulator as the required training approach for physicians who will be using the new device (Gallagher and Cates 2004a, b; Reinhardt-Rutland and Gallagher 1995). The company manufacturing the CAS system informed the FDA that they would educate and train physicians in catheter and wire handling skills with a high fidelity VR simulator using a curriculum based on achieving a level of proficiency. This approach allows for training of physicians to enter with variable knowledge, skill, or experience and to leave with objectively assessed proficient knowledge and skills. This is particularly important for a procedure like CAS as it crosses multiple clinical specialties with each bringing a different skill set to the training table. For example, a vascular surgeon has a thorough cognitive understanding of vascular anatomy and management of carotid disease, but may lack some of the psychomotor technical skills of wire and catheter manipulation. Conversely, an interventional cardiologist may have all of the technical skills, but may not be as familiar with the anatomical and clinical management issues. A sound training strategy must ensure that all of these specialists are able to meet an objectively assessable minimum level of proficiency in all facets of the procedure. This development helps to consolidate the paradigm shift in procedural-based medicine training and will result in a reduction in "turf wars" concerning future credentialing for new procedures. Indeed this was the approach advocated by a number of the professional medical organizations (i.e., vascular surgery, interventional cardiology, and vascular medicine and biology) intimately involved in training physicians for CAS (Rosenfield et al. 2005). As long as a physician is able to demonstrate that he or she possesses the requisite knowledge and skills to perform a procedure, specialty affiliation will become less important. Proficiency-based progression training has leveled the playing field in terms of territorial claims about specific procedures. Decisions about who carries out such procedures will be based firmly on who can perform the procedure to a safe level of skills rather than who has traditionally looked after a particular group of patients. This approach will have profound implications for the practice of medicine. Although we have shown that proficiency-based progression is a better way to train for the *in vivo* practice of procedural medicine, surgical training is about more than just procedural skills. We shall examine this issue further in Chap. 9 when we discuss how we can use the experience and knowledge gained from the development of proficiency-based progression training and augment this approach with e-learning.

Summary

Although medicine in general and surgery in particular profess to be using a competency-based training program, there seems to be no clear operationalized definition of what competency is and what it is not. There has been a considerable amount of effort made by training organizations around the world on competency; however, these efforts have mostly been directed at describing what factors are characteristic of competent performance. Efforts to measure competency appear to have been more comprehensive and systematic in the UK than in the USA. The strongest of the competency assessment procedures in the UK is the procedure-based assessment instruments which have been developed for all index surgical procedures. However, even this instrument could be considerably strengthened with more detailed assessment of the intraoperative performance of the trainee surgeon based on a task analysis as described in Chap. 5.

We have proposed that instead of using competency as the benchmark, it makes more sense to use proficiency as it is not lumbered with the same historical baggage as the concept of "competency" and is easier to establish a widely agreed upon operational definition, i.e., "proficiency is what experienced surgeons (or physicians) do." A proficiency-based training program can be developed using the following steps;

1. Perform the *task analysis* on the procedure to be learned.
2. *Metric definition*: Operationally define the key aspects of optimal procedure performance identified from the task analysis.
3. *Metric validation*: Ensure that metric-based assessment of novice trainee performance differs from experienced operator performance (i.e., construct validity).
4. *Proficiency definition*: Quantitatively assess the performance of a representative number of experienced operators (e.g., consultant/attending surgeons) on the training device/strategy to be used for trainees.
5. *Proficiency-based progression training*: Trainees train on the training device/strategy until they demonstrate the benchmark performance, consistently.
6. *Validate proficiency-based progression training*: Establish whether trainees on the training program perform better than surgeons who were traditionally trained.

The results from preliminary clinical trials using proficiency-based progression training have shown that trainees perform significantly better than traditionally trained surgeons. This approach to training has given further impetus by the FDA in the USA who in 2004 mandated training on a virtual reality simulator for carotid artery stenting. They took this decision in the interest of patient safety to ensure skills of sufficient standard are acquired by surgeons, cardiologists, and radiologists before performing the procedure on patients. Their decision set a precedent which we believe will further drive the changes in training procedural skills in medicine.

References

Ahlberg G, Enochsson L, Gallagher AG, et al. Proficiency-based virtual reality training significantly reduces the error rate for residents during their first 10 laparoscopic cholecystectomies. *Am J Surg*. 2007;193(6):797-804.

Aretz HT. How good is the newly graduated doctor and can we measure it? *Med J Aust*. 2003;178(4):147-147.

Barrows HS, Abrahamson S. The programmed patient: a technique for appraising student performance in clinical neurology. *Acad Med*. 1964;39(8):802.

Beall DP. The ACGME institutional requirements: what residents need to know. *J Am Med Assoc*. 1999;281(24):2352.

Bryant J. *Health and the Developing World*. Ithaca: Cornell University Press; 1969.

Dictionary CC. *Thesaurus*. New York: Harper Collins; 1995.

Dreyfus HL. *What Computers Can't Do: the Limits of Artificial Intelligence*. New York: HarperCollins Publishers; 1979.

Dreyfus HL, Dreyfus SE, Athanasiou T. *Mind over Machine*. New York: Free Press; 1986.

Dreyfus HL, Dreyfus SE, Athanasiou T. *Mind over Machine: The Power of Human Intuition and Expertise in the Era of the Computer*. USA: Simon and Schuster; 2000.

Epstein RM, Hundert EM. Defining and assessing professional competence. *J Am Med Assoc*. 2002;287(2):226.

Frank JR. *The CanMEDS 2005 Physician Competency Framework. Better Standards. Better Physicians. Better Care*. Ottawa: The Royal College of Physicians and Surgeons of Canada; 2005.

Gallagher AG, Cates CU. Approval of virtual reality training for carotid stenting: what this means for procedural-based medicine. *J Am Med Assoc*. 2004a;292(24):3024-3026.

Gallagher AG, Cates CU. Virtual reality training for the operating room and cardiac catheterisation laboratory. *Lancet*. 2004b;364(9444):1538-1540.

Gallagher AG, Satava RM. Virtual reality as a metric for the assessment of laparoscopic psychomotor skills. Learning curves and reliability measures. *Surg Endosc*. 2002;16(12):1746-1752.

Gallagher AG, Richie K, McClure N, McGuigan J. Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. *World J Surg*. 2001;25(11):1478-1483.

General Medical Council. *Tomorrow's Doctors: Recommendations on Undergraduate Medical Education*. London: GMC; 1993.

Grantcharov TP, Kristiansen VB, Bendix J, Bardram L, Rosenberg J, Funch-Jensen P. Randomized clinical trial of virtual reality simulation for laparoscopic skills training. *Br J Surg*. 2004;91(2): 146-150.

Harden RM, Stevenson M, Downie WW, Wilson GM. Assessment of clinical competence using objective structured examination. *Br Med J*. 1975;1(5955):447.

Lurie SJ, Mooney CJ, Lyness JM. Measurement of the general competencies of the accreditation council for graduate medical education: a systematic review. *Acad Med*. 2009;84(3):301.

Marshall JB. Technical proficiency of trainees performing colonoscopy: a learning curve. *Gastrointest Endosc*. 1995;42(4):287-291.

Miller GE. The assessment of clinical skills/competence/performance. *Acad Med*. 1990; 65(9):S63.

Neufeld VR, Maudsley RF, Pickering RJ, et al. Educating future physicians for Ontario. *Acad Med.* 1998;73(11):1133.

Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. *Ann Intern Med*. 2003;138(6):476.

Pitts D, Rowley DI, Marx C, Sher L, Banks T, Murray A. *A Competency Based Curriculum for Specialist Training in Trauma and Orthopaedics*. London: British Orthopaedic Association; 2006.

Popper KR. *Objective Knowledge: An Evolutionary Approach.* Oxford: Clarendon Press; 1979.

Reinhardt-Rutland AH, Gallagher AG. Visual depth perception in minimally invasive surgery. In: Robertson SA, ed. *Contemporary Ergonomics*. London: Taylor & Francis; 1995:531-536.

Rosenfield K, Babb JD, Cates CU, et al. Clinical competence statement on carotid stenting: training and credentialing for carotid stenting–multispecialty consensus recommendations: a report of the SCAI/SVMB/SVS Writing Committee to develop a clinical competence statement on carotid interventions. *J Am Coll Cardiol*. 2005;45(1):165-174.

Seymour NE, Gallagher AG, Roman SA, et al. Virtual reality training improves operating room performance: results of a randomized, double-blinded study. *Ann Surg*. 2002;236(4):458-463; discussion 463-454.

Van Sickle K, Ritter EM, Baghai M, et al. Prospective, randomized, double-blind trial of curriculum-based training for intracorporeal suturing and knot tying. *J Am Coll Surg*. 2008;207(4): 560-568.

Verheugen G. (2005) *Good Clinical Practice.* Retrieved. from http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2005:091:0013:0019:en:PDF. (accessed 10 July 2010).