



**GUIDELINES FOR  
ASSESSMENTS IN NEUROSURGICAL EDUCATION**

June 30, 2007

**Foundation for International Education in Neurological Surgery**

# FOUNDATION FOR INTERNATIONAL EDUCATION IN NEUROLOGICAL SURGERY

## ASSESSMENT in NEUROSURGICAL TRAINING

The following has been prepared by the Board of the Foundation for International Education in Neurological Surgery as a resource to assist cooperating Neurosurgical training Programs and National Societies in the development of assessment processes for resident training and the development of National Certifying Examinations.

It is hoped these guidelines will be helpful to our colleagues. They are based on educational principles of assessment and on practices commonly used by surgical programs throughout the world.

### Preamble:

The primary purposes for assessment are:

1. To provide regular feedback and supervision for performance improvement during training (formative assessment)
2. To determine a satisfactory level of competence at completion of the program (summative assessment).

Assessment is critical to learning and the development of competence. Development of a comprehensive assessment system for specialty training is based on important educational principles (Appendix 1). In any clinical program, assessment is based on applied (rather than simple factual) Knowledge, and clinical performance of desired Competencies. Assessment processes utilize multiple methods to assess multiple competencies, using an adequate and representative sample to ensure reliability and validity. Any comprehensive assessment process will be based on the listed objectives and desired competencies and must include the assessment of the following components:

- Applied Knowledge
- Skills (clinical, technical and procedural)
- Clinical Reasoning and Decision making
- Professional behaviours

**High stakes assessments** are those which permit promotion to a new level of training or those which determine completion and Certification of exit competencies and enable licensure for clinical practice. All high stakes assessments must meet criteria of reliability and validity. **Reliability** is a measure of the reproducibility or consistency of the assessment tool. **Validity** is the ability of a tool to measure what it is intended to measure. All high stakes assessments must be subjected to a rigorous **Standard Setting** process in determination of Pass/Fail criteria (Appendix 1).

## **Comprehensive Assessment Process for Postgraduate Programs:**

Based on sound educational principles and the specific needs of the Program or Society, a system of assessment might include:

### **1. Self-Assessment of Competencies:**

Trainees are encouraged to assess their own progress. Self-Assessment is a case-based system where the trainees critically self-assess their ability to diagnose or manage a case that is encountered in practice. A case is prepared and reviewed with the senior faculty member who provides appropriate feedback. A brief record can be placed in permanent file.

### **2. Periodic In-Training Assessments of trainees:** these may be regular, informal or formal and can include several methods.

1. Formal or Informal in-training assessment using:
  1. *Global clinical assessment forms.* Within a regular time frame (3 months) or at the completion of each component of training, preceptor faculty members complete a standard rating form assessing clinical performance (Appendix 3). These are reviewed with the trainee and permanent records will be kept (In-Training Evaluation Reports – ITER).
  2. *Written examinations.* Regular formal assessments of knowledge (using MCQ or short answer questions – Appendix 2) may be held at the completion of major blocks of training or in accordance with the residency training curriculum program. Trainees should be given the opportunity to complete an examination at the end of each significant component of training which is helpful for self assessment or may be required for promotion to the next level of training. In some programs, formal or informal assessments may be conducted on a periodic basis (6 months).
  3. *Oral examinations.* In some programs, oral examinations may be used to assess clinical reasoning and decision making skills (Appendix 4).
  4. *Observational assessments.* Assessment of clinical or procedural skills by direct observation may be done using the current method of Mini-CEX (Appendix 5). If sufficient numbers of these observations are obtained during the course of the training program, then reliable conclusions can be made about competence and performance.
  5. *Assessment of Error.* Medical errors and issues of patient safety need to be critically assessed in a fair, constructive and formal manner (Appendix 6).

### 3. National Progress Assessment of trainees:

Some Societies are able to provide formal “progress assessments” yearly. The purpose of the examination is to determine progress of knowledge and skill acquisition over the course of training. A written examination is provided to trainees at all levels and results are compared with scores from the previous year’s examinations. Results are made available to trainees and program directors as indicators of knowledge gained.

### 4. National Board Certification Examinations:

Certification Examination Processes present a challenge to many Programs and National Societies. These are **high stakes examinations** which often confer specialty status and determine eligibility to practice neurosurgery in the country. These examinations are usually conducted yearly or as needed (6 months) and are the responsibility of a recognized Examination Board appointed by the National Society. Members of these Examination boards should be knowledgeable about assessment principles and practices.

Because of the importance of this high stakes process, all examinations must meet the rigors of reliability and validity and provide a fair opportunity for the candidate to demonstrate the required knowledge, reasoning skills and clinical competence.

Most National Certification examinations consist of written and oral examinations. Some will also include an Objective Clinical Examination (OSCE).

1. *Written examination* should consist of carefully constructed standardized questions to assess applied knowledge. Good studies have demonstrated that 120 questions will provide good reliability. The questions are prepared following internationally accepted assessment guidelines. (see Appendix 2)
2. *Oral examinations* should consist of a set of standardized cases prepared by the Examination committee. Reliability must be maintained by providing a minimum of 10 cases or topics (ideally this should be 16) and by allowing each candidate to be examined by at least 4 examiners (e.g. 2 or 3 examination teams of 2 examiners). The examination may be 2 – 3 hours in length. (See Appendix 3). Because of serious flaws in oral examination scoring, strict adherence to Standard Setting procedures must be maintained.

3. *The Objective Structured Clinical Examination (OSCE)* is a very reliable assessment of clinical skill however it is costly and requires special expertise to deliver. Methods and guidelines are available in the literature for those programs that wish to use them.

## 5. Standard Setting for Pass/Fail decisions:

All High stakes examinations (especially Certification Examinations) should undergo the rigors of setting Standards for Pass/Fail decisions. Standardized and *agreed upon answers must be prepared* for all components and *pass/fail criteria for each component* must be determined by the team of examiners (a minimum of 2 persons for each component). Satisfactory performance is based on demonstrated applied knowledge, decision making and achievement of required competencies (known as criterion referenced assessment). A Pass/Fail cut-off must be determined by the examination team for the whole examinations process.

There is extensive literature on standard setting procedures.

## 6. Establish a National Examinations Process

In order to provide a high quality examinations process for Certification, societies may consider the following suggestions:

1. Establish a National Examinations Board (Committee) with membership, expertise, responsibility and authority to plan and execute a national examinations system.
2. Establish a format for the Examinations system
  1. Dates, times and venues
  2. Written components
  3. Oral components
  4. Practical clinical components
3. Decide on “examiners”. The Board has the responsibility to appoint an appropriate number (6 - 8) and scope of examiners. Care needs to be exercised so that examiners will be impartial and able to render judgments of candidates based on demonstrated competence. A “chief examiner” can be appointed and is responsible for all aspects of preparation and implementation of the examinations process. It may be desirable to have “external examiners” from other disciplines or from abroad.
4. All examiners should be generally informed with instruction of principles for high stakes examinations.
5. A strict Standard Setting process should be established for each examination.

Further information providing details on the guidelines above can be found in the following Appendices.

Prepared: January 31, 2007

## Appendix 1

### Important Educational Principles of Assessment

Any assessment system within training programs will have several purposes. The primary purposes for assessment are to provide regular feedback and supervision for performance improvement during training (formative) and to determine a satisfactory level of competence at completion of the program (summative). It is particularly important that those assessments which are used for Pass/Fail determination (High Stakes Examinations) meet very strict requirements to ensure high quality and fairness to the candidates and provide reassurance to the public. .

*Case specificity:* every question item in an examination assesses the application of knowledge of the candidate. Each assessment component requires knowledge and skill which is specific to that topic or disease entity and is not generalizable to other entities. Judgment of a candidates overall performance, knowledge or skill cannot be based on a single case. Judgment of a candidate must be made over a wide range of knowledge and skill and therefore must cover a broad sampling of multiple topics.

*Reliability:* is a measure of the reproducibility or consistency of the assessment tool. If a similar examination is given on repeated occasions, reliability means that the results will all be similar. Inter-examiner reliability is important and is enhanced by training of examiners to ensure the same methods and standards of questioning and scoring are used. Case or content reliability is ensured by adequate numbers of topics or questions. It has been shown that 120 questions will provide high reliability scores for written examinations using MCQ or short answer methods. It has also been shown that 16 - 20 topics or themes for oral examinations and 14 – 16 stations for OSCE will provide high reliability scores. Compliance with these guidelines will provide good reliability for high stakes examinations.

*Validity:* is the ability of a tool to measure what it is intended to measure. For example, MCQ are good tools to measure knowledge and application of knowledge. With correct construction, they may also be used to assess problem solving. OSCE examinations are the best tool to measure competence of clinical skills. Oral examinations are good tools to test clinical reasoning and decision making skills but only indirectly assess factual knowledge. Mini CEX is an excellent tool to assess clinical skills, technical, procedural or presentation skills. Consequently, the use of multiple appropriate tools is important to ensure validity of an examination process.

*Sampling:* because knowledge is specific to each clinical case, a broad sample is required for the test to be reliable and the results defensible. All methods should take representative samples from a wide breadth of topics or disease conditions within the discipline. As noted above, an adequate sample number to test knowledge with MCQ or short answer methods is 120 questions whereas the ideal sample number to test decision making by oral examination methods is 16 cases.

*Standard Setting:* establishing standards is critical to high stakes examination processes. These determine Pass/Fail and are usually based on the *minimum* standards acceptable for safe practice of the specialty. Standards are determined by practicing clinicians of the specialty. Although there are different methods for determining standards, they are based on the minimal level of competency required (termed criterion referenced) and a candidate is judged by the ability to demonstrate achievement of that competency.

Standardized *answers* for each question or topic must be prepared prior to the examination and agreed upon by the examiners. The degree of mastery required is determined by the examiners. Standard setting procedures should be established by the examinations committee or “chief examiner” and is a critical part of the examinations process. This should be done at the beginning of each examination setting by the examination team. It is important to determine what *cut-off* is accepted for pass/fail. This should be done by at least 2 persons for each component (question) and the whole examination team should determine the minimum passing score for the examination.

*Scoring:* There are inherent problems of scoring in components involving oral examinations and OSCE methods. Candidate responses must be compared against the key features or standardized answer which has been prepared for each question or topic. Assignment of numeric measures is helpful. Examiners should be provided with instruction to reduce the impact of inherent flaws and bias and to ensure the examination is fair to candidates.

## Appendix 2

### Preparation of written examinations

Multiple Choice Questions (MCQ) are reliable, valid and cost effective methods of preparing examinations. They are primarily used to assess knowledge and are a very popular method for all manners of knowledge assessment. Despite this, MCQ are commonly misused and lead to unwanted consequences. The greatest flaw in the use of MCQ is that they tended only to test simple recall of factual knowledge and hence promote memorization rather than understanding. Current standards now use MCQ to assess application of knowledge or *Applied Knowledge*. This is enabled by careful construction of examination questions. The following guidelines are provided to enable cooperating programs to develop high quality examinations which avoid the problems commonly encountered in the use of MCQ.

Although many types of questions are in use, those recommended are based on the following 2 question types: 1) One best answer questions, 2) Short answer questions

The “One best answer” questions have become the standard method of written assessments. They provide a clear question which can be answered ONLY by the one best answer.

**Basic Rules for One-Best-Answer Items** (Adapted from Case and Swanson). Each item should assess *application of knowledge*, focus on an *important concept*, and be based on a relevant *Clinical Vignette*. A well constructed question consists of a clinical stem, a lead-in question, and a series of answer response options.

Do NOT write True/False type questions.

Do NOT write questions in the form - “Which of the following statements is correct?”

Do NOT write negative stems – e.g. “Each of the following statements is correct EXCEPT.”

Do NOT write questions / answers which provide multiple options: a, b and c; a and c; none of the above, all of the above etc.

#### References:

- Case and Swanson, NBME “Constructing Questions for the Basic and Clinical Sciences”  
<http://www.nbme.org/about/itemwriting.asp>
- Royal College of Physicians and Surgeons of Canada website  
<http://www.rcpsc.medical.org>

## Practical Steps in constructing a question

### A. Choose a topic or theme

The topic is the specific topic or medical knowledge that is to be tested. Focus on a single important concept, typically a common or potentially serious clinical problem from every day practice. Some examples might be: intracranial neoplasia, increased intracranial pressure, spinal trauma, subarachnoid hemorrhage. Don't waste time with questions assessing knowledge of trivial facts.

For fairness, test topics must be based directly on published objectives, defined competencies or desired outcomes. Examination committees should provide guidance to examiners who prepare questions on what content is appropriate.

### B. Choose the appropriate clinical context

Context refers to a typical clinical situation that requires knowledge or use of the topic. An example might be from an emergency department which admits trauma or acute intracranial emergencies, the ambulatory clinic or a ward.

### C. Create the stem

#### 1. Construct a clinical case vignette

Clinical cases provide a good basis for a stem. The clinical case should be relevant and related to problems that would be encountered in an average clinical practice rather than trivial, or complicated obscure problems. Begin the scenario by presenting a problem followed by relevant signs, symptoms, results of diagnostic studies, treatment, subsequent course etc. In essence, provide all essential information that is needed for a competent candidate to answer the question.

#### 2. Test the application of knowledge

The question should test *applied knowledge* rather than just the simple recall of factual information. This requires understanding of knowledge and how it applies to clinical situation rather than recall of memorized facts.

### 3. **Test an appropriate level of difficulty**

Well constructed questions should be designed to test the knowledge appropriate to the level of the candidate. For example, Certification examinations test knowledge required to begin the first day of clinical practice in the specialty.

#### **D. Ask a clear question**

The question must be clear and require the candidate to consider the knowledge and information needed to provide an answer applied in the context of the scenario. The correct answer presumes the required knowledge. Do not ask a question which can be answered by simple fact recall of knowledge. (Refer to Case and Swanson)

#### **E. Write the correct answer**

The correct answer should be clearly correct and should be the single best answer for the question. The answer should be obvious to the candidate who has appropriate knowledge.

#### **F. Create the distracters**

Three or four distracters should be provided. They should be homogeneous in content, the general same length as the correct answer, consistent in grammar and syntax. A good distracter may be plausible but should clearly be inferior to the correct answer. Avoid overlap in the content of distracter options. Avoid ambiguous and confusing terms such as 'frequently, often, sometimes' etc.

#### **G. Avoid technical item flaws** such as grammatical or syntax clues that provide special benefit to test-wise examinees or that pose irrelevant difficulty.

(Examples can be found in the reference material of Case and Swanson)

## Appendix 3

### Global Clinical Performance Assessment

Global Clinical Performance Assessments are the traditional “in-training” assessment forms (ITER) used in all programs. The purpose is to provide encouragement to resident performance and feedback for improvement.

Usual guidelines recommend these be completed and reviewed with every resident at the end of each block of training or a minimum of every 6 months. Honest assessment by each faculty member is helpful and a composite prepared by the program director for discussion with the resident.

There are many variations found in the literature but “simple is best”. Usually four categories are assessed:

Applied Knowledge

Clinical Skills including Decision Making skills

Procedural Skills

Professional Behaviours

*Rating forms* should be simple and easy to fill out. Sub headings under each category should be of limited number, use simple language and assess the significant and important aspects of performance. Items under each should be specific, understandable and assessable by both staff and residents. Each one should assess a key component of performance. Many rating forms fail because they are too complicated.

*Rating scales* provide an assessment over a range of performances. Most have been based on principles of a Likert scale or a range (1 – 7). However it is found that a simple 4-point performance rating is more descriptive, offers a judgment on performance and provides opportunity for effective feedback and planning. These use descriptive language such as:

1. Unsatisfactory; 2. Needs Improvement; 3. Meets Expectations; 4. Exceeds expectations

Most trainees will perform at “Meets Expectations”. “Needs Improvement” identifies a level of achievement which requires action but without indicating failure. Good rating forms will provide brief descriptors of expected performance for each level.

## Appendix 4

### Preparation for Oral Examinations

Oral examinations are a common method of assessment but are fraught with many problems. In order to ensure high quality oral examinations, several important principles must be followed.

Fairness: examinations must be fair to the candidates and provide them opportunity to demonstrate their knowledge and skills over a wide range of disease conditions. This requires sufficient sampling from a breadth of topics.

Validity: oral examinations are not a good method to assess knowledge. This can be done much better through MCQ or short answer questions. They are however, good ways to assess Clinical Reasoning, Decision making and Clinical Judgment. Consequently questions focus on problem solving and decision making around common clinical themes or disease conditions.

Reliability: consistent with the principle of sampling, good studies have shown that 20 themes of topics will provide good reliability for the examination. A balance must be addressed between the ideal and what is reasonable and feasible, however examinations committees should strive for 16 themes or questions to ensure reliability. This will require careful planning and strict time management of the examination process.

A practical plan could be to provide three or four “stations” of 2-3 examiners, each of which can address 4 - 5 topics in a one hour period.

Question construction: question preparation will follow principles similar to those for written questions, where a case scenario is prepared which presents a relevant clinical problem followed by an open question. The topics should be based in common disease conditions or those of a serious nature. Complex conditions from subspecialty interests, obscure or rare conditions should be avoided. The questions are asked in a manner so as to elicit problem solving or decision analysis rather than recount factual information.

Determination of which questions are used for the examination can be decided by the board of examiners prior to the exam.

Answer preparation: to ensure consistent and reliable scoring for the question, a prepared “ideal answer” consisting of the required “key features” is essential. At the pre-exam meeting, a model answer of “key features” can be agreed upon by examiners. Scoring of candidate performance is based on identification of key features in the answer. Particular attention is paid to organization and attention to development of diagnostic reasoning and therapeutic reasoning for problem management.

Standard setting: It is very important that the examination committee and the examiners meet prior to the examination to determine ideal answers for each question, and to determine minimum competency requirements and to determine the cut-off standard for Pass / Fail.

Training of examiners: to ensure high quality and reliability of the examinations, an instructional program for examiners could consist of the concepts of reliability and validity, construction of examination questions, methods of scoring, and standard setting for pass / fail determination.

## Appendix 5

### Observational Clinical Examination Exercises (Mini CEX)

The Mini CEX is a new development for assessment of performance. It has many advantages in the training program to trainees and staff. It is an *observational assessment* usually involving 10 – 20 minutes to complete the following steps:

1. Observation of a clinical performance,
2. Reference against criteria of expected performance (checklist)
3. Immediate feedback to the trainee.

The advantages of Mini CEX are:

- a. they can be done within the context of regular practice and teaching
- b. they add minimal increased time to teaching
- c. they are flexible and can be applied to almost any aspect of training including: history taking, physical examination skills, communications, case presentations, reasoning and decision making and technical or procedural skills.
- d. they provide immediate feedback to the trainee on performance
- e. Reliability and Generalizability is achieved with 8 – 10 assessments

Although the primary purpose is to provide formative improvement, if done over an extended period of time, they can sometimes be used as summative assessments with pass/fail consequences.

The challenges to implementation are the development of specific checklists for expected performances in the variety of venues where they can be used.

*Implementation:* Identify what venues are suitable for Mini CEX assessment. These could include: aspects of data gathering (such as history taking, physical examination, interpretation of imaging or laboratory investigations); reasoning and decision making skills; technical or procedural skills and case presentations.

Prepare simple check lists of expected performance for each of the several venues where they will be used.

Ensure that feedback is provided with each assessment.

Develop a schedule of assessments (perhaps 6 per year over the course of training). There can be shared trainee and preceptor responsibility in identifying and initiating opportunities for these assessments.

## Appendix 6

### Assessment of Error and Adverse Patient Outcomes (APO)

Critical analysis of Adverse Patient Outcomes (APO) is an important part of program development and learning in patient care. It is critically important that this be conducted in the most professional manner so that issues of care are addressed and learning is a priority. Discussion of patient safety issues and critical reviews of APO's must be effectively chaired and managed in a safe constructive environment rather than one of criticism, personal attack or punishment. The purpose of all discussion is for learning and improvement in patient care outcomes. Although there are several frameworks for analysis, the following is proposed as functional and thorough.

**Type of APO:** Adverse Patient Outcomes are usually the result of the natural course of the disease, an organizational System failure or due to Human error. Although most APO are due to System factors within health care institutions, this framework will address issues of human error in the context of surgical practice and training.

**Analysis of Adverse Patient Outcomes** should focus on several factors.

*Problem of the disease:* is the adverse outcome the result of the severity or problem directly caused by the disease process?

*Diagnosis:* has the problem arisen due to an error in diagnosis?

*Knowledge:* has the problem arisen because of a deficiency in knowledge?

*Reasoning, Judgment and the Decision process:* has the problem arisen because of incomplete reasoning or errors in decision and problem solving?

*Procedure planning:* has the problem arisen due to inadequate or incorrect planning?

*Technical or procedural:* has the problem arisen due to an intraoperative technical problem or an incorrect step in the procedure?

*Intraoperative Cognition or Decision making:* once the problem arose, was there an error in intraoperative response or decision making?

Discussions should ideally occur with all staff and trainees present with the focus addressing the following issues:

What can we learn from this problem?

What can we do differently?

What can be done to prevent similar problems in the future?

**“Near Miss” analysis:** We can learn from the aviation industry which has pioneered these preventative analyses. Recent literature on medical error and current thinking is now encouraging us to address *incidents of “near miss”* – i.e., where an error or problem arose which did not cause an adverse outcome for the patient. Whereas in the past, these were considered ‘lucky escapes’ and not analyzed, they are now viewed as wonderful opportunities for learning and improvement in patient care. Because they are free of serious consequence, they can be discussed candidly and true preventative actions can be instituted to prevent recurrence.

## References:

### Principles:

Downing, SM. Validity: on the meaningful interpretation of assessment data. *Medical Education* 2003; 138: 476-481

Downing, SM. Reliability: on the reproducibility of assessment data. *Medical Education* 2004; 38: 1006-1012

### National Certification Examinations:

Royal College of Physicians and Surgeons of Canada.  
<http://www.rcpsc.medical.org>; - publications and documents-material for examiners

Accreditation Council for Graduate Medical Education. <http://www.acgme.org>

### Oral Examinations:

Wakeford, R., Southgate, L. & Wass, V. (1995). Improving oral examinations: selecting, training, and monitoring examiners for the MRCGP. *BMJ* 1995; 311:931-935 (7 October).

Wass, V et al. (2003). Achieving acceptable reliability in oral examinations. *Medical Education* 2003; 37: 126-131.

Yaphe, J. & Street, S. (2003). How do examiners decide? a qualitative study of the process of decision making in the oral examination component of the MRCGP examination. *Medical Education* 2003; 37: 764-771.

Thomas, C. S. et al. (1992). The oral examination: a study of academic and non-academic factors. *Medical Education* 1992; 27: 433-439.

### Written Multiple Choice Questions:

Case and Swanson, NBME "Constructing Questions for the Basic and Clinical Sciences"  
<http://www.nbme.org/about/itemwriting.asp> - publications, item writing manual.

Royal College of Physicians and Surgeons of Canada  
<http://www.rcpsc.medical.org> - publications and documents-material  
for examiners

Standard setting:

Norcinii, JJ. Setting Standards on Educational Tests. *Medical Education* 2003; 37: 464-469

George, S et al. Standard setting: comparison of two methods. *BMC Medical Education*, 2006; 634-636. <http://www.biomedcentral.com/1472-6920/6/46>

Searle, J. Defining Competency – the role of standard setting. *Medical Education* 2000; 34: 363-366

Cusimano, M. Standard Setting in Medical Education. *Academic Medicine* 1996; 71: 112-120

Downing, SM et al. Procedures for establishing defensible absolute passing scores on performance examinations in health professions education. *Teaching and Learning in Medicine* 2006; 18:50-57

Mini CEX:

Norcinii, JJ et al. The Mini-CEX: a method of assessing clinical skills. *Ann Intern Med* 2003; 138: 476-481.

Hatala, R et al. Assessing the mini-Clinical Evaluation Exercise in comparison to a national specialty examination. *Medical Education* 2006; 40: 950-956.