TASK FORCE ON EDUCATION IN EUROPEAN PSYCHIATRY

APPENDIX I

Glossary of terms

The *national/regional curriculum* within a particular country/region describes and contains

a *syllabus* with the elements of theoretical knowledge and their description
a list of *basic rotations, episodes of general psychiatric training within a particular clinical setting* each trainee should have participated in
a list of *subspecialty rotations* a trainee might choose from after finishing the basic training
a list of *skills* a trainee should acquire
the *necessary competencies for trainers*
the *tools for evaluation* of trainees as well as trainers

A *training institution* is recognised by the competent authority within a particular country/region

A *training program*

is the set of training elements offered by an established training institution

it can be *complete* if this program allows for being recognised in a specific field of competency after successful finishing it

it can be *partial* if it constitutes only a recognised part of the training

A trainee can only be a *recognized specialist in psychiatry* after graduation from their training institution

Credentials are documents that are proof of the acquired competencies and delivered by the competent authority in their country/region

A *license to practice* can be obtained from the competent authority within a country/region and allows a recognised specialist in psychiatry to practice in a particular country/region

Fulfilling criteria concerning *continuous medical education* might influence the duration of validity of this licence to practice

APPENDIX II

We refer to the European Framework of Competencies in Psychiatry which we take the following paragraphs out for reasons of congruency.
(References in the text can be found in the original document on the web)

Introduction

This glossary accompanies the assessment grid of the European Framework for Competencies in Psychiatry (EFCP). The assessment grid shows suggested methods of assessing the competencies. The purpose of this glossary is to explain what the different methods are and to give examples of how different tools based upon these methods may be used in practice.

It is now widely recognized that assessment drives learning therefore an assessment system must be considered as being an integral part of any curriculum that is to be developed from the competency framework. This applies as much to professional training as to continuing professional development.

There are three principles that should guide the construction of assessment systems:

- Assessment systems should be transparent, so that learners and teachers know what is being assessed and how it will be assessed.
- Each competency should be assessed, not just those that are easy to assess
- Competency assessment must be triangulated, that is each competency must be assessed in more than one way on more than one occasion.

A further consideration is the utility of the assessment system. Van der Vleuten (1996) pointed out that in mathematical terms the utility of an assessment system might be considered as the product of its reliability, validity, feasibility and educational impact (that is the effect that assessment has upon learning). It follows that if the value any of these qualities approaches zero, no matter how positive the remaining values are, the utility of the assessment system will also approach zero.

Miller (1990) described a conceptual model of the different domains of medical skill and how they may be assessed. Miller's model emphasizes the importance of the assessment of performance (that is, what the doctor actually does in their day-to-day practice), rather than surrogates, which are actually assessments of knowledge or competence.

In this assessment grid, we have sought to identify at least two methods of assessment for each competency. For ease of viewing, we have arranged the assessment methods into one of the three domains in Miller's model, knowledge ('what the doctor knows'), competency ('what the doctor can do') and performance ('what the doctor does'). In the following sections of the glossary, we will describe each method of assessment and what is known about the reliability, utility, feasibility and educational impact of tools that are based on the methods, so that national associations and other regulators of psychiatric training may make informed choices regarding assessment methods. We will give more attention to the tests of the 'does' level, as they are likely to be less familiar to readers.

KNOWLEDGE ASSESSMENTS (TESTS)

Written examinations (WE)

There are two main types of written assessment: multiple-choice papers, in which the candidate selects the correct response from a number of alternatives and essay papers or short answer papers, in which the candidate has to construct text.

Multiple-choice questions: Papers based on multiple-choice questions (MCQ) offer a high degree of reliability per hour of testing time (Schurwirth and van der Vleuten, 2003) and if constructed well, they can test more than factual recall. There are now several question types available in addition to the traditional 'true/false' format. They clearly offer a reliable, valid form of assessment as long as due care is given to the construction and evaluation of questions. The facility to mark MCQ's electronically contributes to their high feasibility.

Essays and Short Answer Papers: Essay papers have been used to examine the ability of candidates to express themselves in writing and to use other intellectual skills (Schurwirth and van der Vleuten, 2003). Indeed, there is a great degree of face validity to this form of assessment in a highly language dependant discipline such as psychiatry. The use of this form of assessment is limited by the time taken to answer essays and hence essays have only limited feasibility. Short answer papers appear to assess similar domains of knowledge as MCQ papers, and since they depend on human markers, they can be less reliable and are also less feasible.

Oral examinations (OE)

Oral examinations may be defined as "examiner/examinee encounters where topics unrelated to specific patients are discussed" (Wass et al, 2003). This form of assessment is intended to assess clinical reasoning and decision-making skills and professional values. Swanson et al (1995) estimated that approximately eight hours of examiner time (either as paired examiners or individual examiner) is needed to produce an acceptable degree of reliability. A similar study of UK general practice candidates indicated that a well structured oral examination covering between 20 and 25 topics over three to three and a half hours of testing could produce acceptable reliability (Wass et al, 2003). The validity of this form of assessment must be carefully monitored, however, as Roberts et al (2000) found evidence the oral examination has a particular potential for bias against candidates from minority ethnic groups.

COMPETENCY ASSESSMENTS

Clinical examinations (CE)

The long case examination is one of the most venerable forms of assessment in medical education (Jolly and Grant et al, 1997). In the long case, candidates are given up to an hour to assess a non-standardised patient. They are assessed on the subsequent presentation they deliver to the examiner(s) and sometimes also on a brief observed interview with the patient. The examination may take up to an hour and a half.

There are serious concerns about the reliability of the long case examination (Jolly and Grant, 1997) and these concerns arise because the assessment is based upon an encounter with one patient and unstructured questioning by examiners (Fitch et al, 2008). Norcini (2002) has reported reliability estimates for a single long case of 0.24. Having more assessments performed by more assessors and observing the whole encounter between candidate and patient increase the reliability of the long case. Six such long case assessments are needed to bring a reliability coefficient of 0.8. Unfortunately, however, the large amount of assessment time needed and the lack of willing and suitable patients severely limits the feasibility of the long case examination.

Assessment of simulated clinical encounter (ASCE)

The ASCE examination seeks to assess clinical competency by rotating each candidate around a number of standardized situations. Typically, each 'station' (encounter) in the examination will consist of a clinical scenario enacted by a role player and the candidate is given a task. The examiner observes the candidate performing the task and marks the performance against a given set of criteria, which is why this form of assessment is widely referred to as the ***Observed Structured Clinical Examination (OSCE)***. Newble and Swanson (1998) found that acceptable levels of reliability are attained after about 16 OSCE stations with one examiner at each station. This equates to about three hours of test time per candidate. The OSCE examination in UK postgraduate psychiatry has been shown to produce similar reliability estimates (Lunn, personal communication). Recruiting and training examiners and role players, as well as finding suitable examination venues, are the factors that most restrict the feasibility of this assessment tool.

PERFORMANCE ASSESSMENTS

This form of assessment is often referred to as ***workplace-based assessment (WPBA)*** to emphasize that it is based upon a doctor's real-time day-to-day work and to distinguish it from standardized tests that may be conducted at a national level or will involve visiting an examination centre away from the place of work.

Fitch et al (2008) identified three methodologies to WPBA:

- The observation and assessment of a doctor's performance conducting their work - direct observation of practice
- The collation of standardized data from several assessors – multi source feedback
- Retrospective assessment of performance through conversations based upon written material, such as log books or clinical records – document-based discussion

To date, very little work has been done evaluating the utility of WPBA in psychiatry; an early field trial in the UK indicated that a programme of assessment based on the three main methodologies outlined above was feasible and acceptable to doctors and their assessors and had some positive educational impact (Brittlebank, 2007). All of the reliability and validity data of the methods has come from areas of medical practice outside psychiatry.

Directly observed practice (DOP)

The DOP method entails an assessor watching a doctor conducting a task, which may involve interacting with a patient, performing a practical procedure or performing a non-clinical task, such as teaching or giving expert testimony. A large number of different DOP tools have been evaluated.

The mini-Clinical Evaluation Exercise (mini-CEX) involves an assessor observing a doctor performing a task, such as history-taking or gaining informed consent, which involves communicating with a patient. It takes around 20 minutes, followed by 5-10 minutes for feedback. The mini-CEX has a large evidence base, with a generalisability coefficient (reliability score) of 0.77 for 8 assessments (Kogan et al, 2003) and reasonable construct validity (Holmboe et al, 2003).

The ***Clinical Evaluation Exercise (CEX)*** involves an assessor observing the doctor conducting an entire clinical encounter with a patient, in this way it is a WPBA equivalent of the long case assessment and it has strong face validity in psychiatry (Brittlebank, 2007). A CEX takes over an hour to perform. Its reliability is quite low; Norcini (2002) reported that two CEX assessments conducted in internal medicine produced a combined reliability coefficient of 0.39.

The Direct Observation of Procedural Skills (DOPS) was developed as a tool to assess a trainee's performance of practical procedures, such as venepuncture or intubation (Wilkinson et al, 2003). Early psychometric data on the DOPS suggests that the reliability and validity of this instrument compares favourably with the data for the mini-CEX (Wilkinson et al, 2008).

The feasibility of DOP-based assessments in psychiatry is determined by the length of time involved in the process, the acceptability to patients of having an observer present in the consultation and (especially in the case of mini-CEX and DOPS) how easily psychiatric practice may be broken down into smaller portions. It is also influenced by the training needed to complete assessments; Holmboe et al (2004) has demonstrated that assessors need to be trained in order for them to be able to conduct fair assessments.

A number of other DOP type instruments are undergoing evaluation; these include tools to assess performance in teaching (Assessment of Teaching), presentation skills (Journal Club Presentation and Case Presentation) and performance of non-clinical skills (Direct Observation of non-Clinical Skills).

Multi-source assessment of performance (MSAP)

MSAP entails the assessment of a doctor's performance from several viewpoints, using a standardized measure that is then collated and fed back to the doctor. The feedback may be from colleagues, both peers and coworkers from different professions and/or levels in the organisational hierarchy, and from patients. MSF may also involve an element of self-assessment.

MSAP has been widely used in professions outside healthcare for many years, where it is more commonly referred to as multi-source feedback or 3600 appraisal (Fletcher, 2004). According to Malik et al (2008) the use of MSAP in medicine has three main attractions:

- Assessments from multiple sources may be perceived as being fairer than assessment from a single source
- MSAP may facilitate assessment of areas of performance (such as the humanistic and interpersonal aspects of medicine) that are not easily assessed using other methods
- To address wider social issues about the accountability of the medical profession.

The feasibility of MSAP is influenced by the availability of competent raters and their access to components of the doctor's practice; raters can only assess that which they can observe and are competent to assess. There will be aspects of practice that peers have not observed and areas that coworkers and patients may not be qualified to comment upon. Feasibility also depends upon the time taken to complete MSAP tools and the ability of the person who collates the data to give helpful feedback to the doctor. Wilkinson et al (2008) have demonstrated that it takes an average of six minutes to complete a typical MSAP form used in medical practice.

The published data on the peer and coworker MSAP tools that have been used in medical training suggest that responses from as few as four (Archer et al, 2006) to 12 assessors (Wilkinson et al, 2008) can produce reliable data. Furthermore, one form, the Sheffield Peer Review Assessment Tool (SPRAT) has been shown to have good feasibility and construct validity data (Archer et al, 2005). A high level of reliability was also demonstrated for nine responses on an MSAP tool (the Team Assessment of Behaviour) that was developed to be mainly a screening tool to identify trainees in difficulties (Whitehouse et al, 2007).

Although a number of tools have been developed to enable patients to give feedback on the performance of their doctor, none has been developed to be used on doctors in training and only two, the Physician Achievement Review (PAR) and SHEFFPAT, have been subjected to reasonably rigorous reliability and feasibility studies (Chisholm and Askham, 2006). These studies indicated that around 25 patient responses were needed to provide reliable data on doctors' performance (Crossley et al, 2005, Violato et al, 2003).

Document-based discussion (DBD)

In this method, a doctor's documented performance in clinical work is assessed through a discussion led by an assessor. There are two main methods in this, discussions based on logbooks or based on patient case records. Although logbooks have been in use in medical training for some time, there is little information in the literature concerning their use as part of a structured assessment. There are several descriptions and evaluations of the use of case records as the focus of assessed discussions – 'Chart Stimulated Recall' (CSR) in the United States. A review of these studies (Fitch et al, 2008) showed that CSR displayed good reliability and validity in assessing medical undergraduates and physicians.

In the CSR, a doctor presents a number of case records to an assessor, who chooses one record to be the focus of the discussion. The assessor questions the doctor on their performance and handling of the case, based on information the doctor has recorded. The discussion allows the doctor to explain their decision-making and can allow exploration of the doctor's clinical reasoning, including the medical, ethical and legal aspects.

The process takes between 20 and 30 minutes to complete and assessors need little training in this method, other than guidance regarding the format of the assessment. It is therefore potentially a highly feasible form of assessment.