# SESSION III:

## SCIENTIFIC THINKING IN MEDICAL EDUCATION - EVALUATIVE AND OTHER ASPECTS -

Chairmen: Prof. N.-H. Areskog (Sweden); Dr. O. Harlem (Norway)

### Clinical Competence: Definition and Assessment

Prof. D. I. Newble (Australia)

Nine years ago, the AMEE Conference was held in Nijmegen. The theme of the conference was "Assessment of Competence in Undergraduate Medical Education". At this meeting I was invited to give two papers - one dealing with the definition of competence and one dealing with the evaluation of competence (Newble, 1981; Newble, 1981). It, therefore, seemed logical to review what I had to say at that time and see what changes had occurred between 1980 and 1989.

With regard to the definition of competence it is disappointing to report that very little new information has appeared over the last 10 years. In their book "Assessing Clinical Competence", Neufeld and Norman include a chapter reviewing the methods used to define competence. The book was published in 1985 yet the most recent reference they quoted relating to definition of competence was published in 1979 and many of the most pertinent were written in the 1960's and early 1970's. Perhaps one might conclude that the problem had been solved and that we have a valid definition of competence. However, this was not the opinion reached by Neufeld & Norman. They concluded that "No single method can adequately define the pre-requisite knowledge, skills and attitudes required of a competent physician" and that the methods used in the past all had limitations derived from bias or too narrow a focus.

The problem is, if we are going to approach the assessment of competence in a way which has any resemblance to the scientific method then we must have a detailed definition on which to base the development of our test procedures and against which we can judge their validity. The definition will determine the objectives of the assessment. The definition will also of necessity be complex and will be composed of a wide range of attributes.

One such definition, which in my view has not yet been improved upon, at least as one appropriate for undergraduate education, arose from a major critical incident study conducted by the NBME in the United States in the 1960's (Hubbard et al, 1965). This study produced a list of nine competence categories (History; Physical Examination; Tests & Procedures; Diagnostic Acumen; Treatment; Judgement and Skill in Implementing Care; Continuing Care; Physician/Patient Relation; Responsibilities as a Physician) each of which was broken down into subcategories. For instance if we took the category Physician/Patient Relation there are three sub-categories (Establishing rapport; Relieving tensions; Improving co-operation). Once again each of these was further divided to provide descriptive statements of the types of behaviour by which each subcategory would be recognised.

Assuming we have some definition of competence available to us when we set about designing our assessment, how are we going to approach detailed definition of content and the selection of test methods in a rational way? Unfortunately, there is often not a clear and logical link between these elements. Assessments are often related more directly to imperatives imposed by departments, disciplines or external agencies. Departments may, for example, lock themselves into using certain methods on the basis of tradition or expediency even though they may be inappropriate. In order to overcome this problem in our own university, Clinical Competence has now become a subject in its own right.

The model on which the approach we have adopted is based has been described elsewhere (Newble, Elmslie & Baxter, 1978). In essence, we link the selection and development of test methods through clinical problems. For each problem we produce a blueprint which identifies

the content we wish to test. The blueprint starts simply as a piece of paper listing the nine competence categories mentioned previously. For each of these subjects specialists are asked to list those key items which students should know or be able to do if they were to successfully deal with that particular problem at the level of competence expected of an intern.

So, for example, if we took the problem chest pain and looked at category three, Tests & Procedures, the key items would include aspects of electrocardiography, chest radiology, cardiac enzymes, coronary angiography and so on. Thus, we are using the defined categories as a check-list to ensure that the content on which we base the examination covers the full range of knowledge and skills over which we expect our students to be competent. It is on such a structured approach that we will have to rely if we are going to establish the content validity of our assessment and it is not possible to overemphasize the fundamental importance of content validity if we wish to produce a good test.

Perhaps I am overcomplicating things so let me try and develop a model which may simplify the message and at the same time introduce a couple of new concepts which need to be included in any discussion on the definition of competence. In recent years there has been a trend to restrict the term competence to the capacity or ability of the student or doctor to do something and seperate it conceptually from "performance" in practice. At its most simple, we might view competence as the mastery of both a body of relevant knowledge and a range of relevant skills (which would include clinical, interpersonal and technical skills). Knowledge and skills are, of course, interrelated but ultimately only useful if they are put to some purpose which we might call clinical problem-solving. Finally, it is probably wise to indicate in our model that competence is only a prerequisite to performance in the real clinical world. Unfortunately, we know from studies in the quality assurance area that competence does not always correlate very highly with performance in practice.

On the basis of such studies, we might argue quite persuasively that clinical assessment should only be based on measures of performance or outcome of patient care rather than on measures of competence. This is certainly a valid argument in the postgraduate period where doctors have direct responsibility for patient care. However, such opportunities are limited in the undergraduate period where we have no option but to look predominantly at competence. This is, perhaps, fortunate as measuring the outcomes of patient care is notoriously difficult.

However we decide to define competence, when it comes to assessing it we must have some kind of matrix which allows us to match the categories of competence with the test methods available. We must, as I have mentioned previously, sample across the full range of problems with which the student must deal. To do this effectively it is necessary for those responsible for assessment to have an understanding of the strengths and weaknesses of the available test methods. Some help is available in this regard. The two best resources are Neufeld & Norman's book and the report of the 1st Cambridge Conference which is entitled "Directions in Clinical Assessment" (Wakeford, 1985). Table 1 is taken from this report and gives a consensus view on the relative merits of various test methods for assessing the different components of competence. Other useful resources are the proceedings of the first two Ottawa Conferences (Hart, Warden & Walton, 1985; Hart & Harden, 1987) and, in due course, those from the third conference held recently in Groningen.

## TABLE 1

**Recommendations on the use of evaluation methods to access domains of competence**

\+ = of some use  +++ = of most use

| Competence/skill | Method | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Global Ratings | MCQ | MEQ | PMP | "Cambridge Case" | Standar-dised Patient | Patient Rating | Direkt Obser-vation | Mechanical simulation |
| 1. Knowledge | | ++ | ++ | + | | + | | + | |
| 2. Interviewing/ Interpersonal | | | | | | ++ | ++ | ++ | |
| 3. Data gathering, History | | | + | + | +++ | +++ | | ++ | |
| 4. Physical Exam. (Technical) | | | | | | +++ | | + | + |
| 5. Reasoning/ Diagnosis | | + | + | + | ++ | + | | + | |
| 6. Lab Utilis./ Management | | + | + | + | ++ | | | | |
| 7. Personal Qualities | ++ | | | | | | | | |

From Directions in Clinical Assessment (1985) Wakeford (Ed)

## TABLE 2

Projected reliabilities at various test lengths (estimated from pooled 1983-86 data). All entries in the table are generalizability coefficients (intraclass correlations) including inter-item/station and interrater (for patient stations) sources of measurement error, but excluding item/station difficulty. Italicized entries indicate the reliability at the test length actually used in the 1985 and 1986 test administrations. Estimated variance components on which reliability calculations were based are available from the second author.

| Test length (hours) | MCQ in medicine | Short answer | Patient stations* | Static stations | Clinical test* |
|---|---|---|---|---|---|
| 0.5 | 0.62 | 0.42 | ***0.31*** | 0.23 | 0.32 |
| 1.0 | ***0.76*** | 0.59 | 0.47 | ***0.38*** | 0.48 |
| 1.5 | 0.83 | ***0.68*** | 0.57 | 0.47 | 0.58 |
| 2.0 | 0.87 | 0.74 | 0.64 | 0.55 | 0.65 |
| 3.0 | 0.91 | 0.81 | 0.73 | 0.64 | ***0.73*** |
| 4.0 | 0.93 | 0.85 | 0.78 | 0.71 | 0.79 |
| 6.0 | 0.95 | 0.90 | 0.84 | 0.78 | 0.85 |
| 8.0 | 0.96 | 0.92 | 0.87 | 0.83 | 0.88 |
| Items/stations per hour | 75 | 44 | 10 | 10 | 22/5 |

* Two raters per patient station.
(Taken from Newble & Swanson, 1988)

To help us move from the rather theoretical approach I have taken so far to a more practical viewpoint, I want to discuss briefly some of the work we have been doing at the University of Adelaide over the last 10 years or so (Newble, 1988). This is simply to provide a case study which will illustrate some of the problems one faces when trying to translate educational theory into practice.

As I mentioned previously we administer a test of clinical competence to all students of the end of the final year. This test is run jointly by the Departments of Medicine, Surgery, Paediatrics and Obstetrics & Gynaecology. It is composed of two equal components of 90 minutes. One is a test of relevant knowledge composed of short answer questions. The other is a structured clinical examination of 15 stations.

Over several years this examination has been subject to a rigorous psychometric analysis by my colleague David Swanson from the NBME in Philadelphia (Newble & Swanson, 1988). As with any critical evaluation the results were not always as we had expected, nor were the messages always the ones we wanted to hear. However, this is the nature of research. Though time precludes any detailed discussion of this work let me highlight a few issues.

We were, of course, interested in providing evidence for the reliability and validity of the examination. I have given you some information about our approach to content validity and I could provide a little evidence on its construct validity (Newble, Hoare & Elmslie, 1981). However, I will restrict my remarks to the issue of reliability.

In table 2, you can see real (&INI.) and projected reliabilities obtained or estimated from pooled data collected over a 4 year period. The estimates come from a statistical approach based on generalisability theory. A clear understanding of the importance of this information can best be obtained by reading the paper from which it is taken (Newble & Swanson, 1988). Nevertheless, you can appreciate the very low reliabilities for most components of the examination except for the written components. The projected reliabilities allow us to estimate how much more testing time would be required for each subsection of the test to achieve satisfactory reliability. The only comfort we can draw from this study is that the reliability for the overall clinical test is acceptable.

We might now ask: What is the reason for the low reliabilities of the clinical components of the test? Traditionally, of course, the main concern with clinical examinations has been with rater reliability. However, in this stuctured clinical examination this does not appear to be the major factor. Average inter-rater reliability works out to be about 0.7. This is about the same as has been found in a number of other studies using a similar approach (van der Vleuten & Swanson, 1989).

The real problem emerges when we look at interstation correlations. They are very low being of the order of only 0.1. There is thus considerably more variance in performance of candidates between stations than there is between marks awarded by raters. This is not an inherent problem with the technique we are using or an Australian aberration but a problem which seems to affect <u>all</u> methods used to assess clinical skills and clinical problem solving.

Clearly there must be a common problem in all these situations and this appears to be "case specificity". This simply means that the performance of a candidate in one clinical situation is not a very good predictor of performance in another clinical situation. Of course, this is not surprising when we take into account recent research into clinical problem solving. We now know that the quality of problem-solving is determined more by specific knowledge and experience with each particular problem than it is by any general problem-solving skill (Norman, 1988).

The difficulty we find ourselves in is that most forms of clinical assessment are based on a very limited number of observations. In the traditional clinical examination used in many parts of the world for assessing students and postgraduates, decisions may be made on the basis of performance on one long case and a handful of short cases. This is clearly an inadequate sample of performance: the evidence is available to prove it, yet such examinations are still widely used. We have to face up to the uncomfortable fact that to achieve a valid and reliable assessment of competence we will have to sample from a large number of clinical problems and across the full range of competence categories. We will also need to use a number of test methods, selecting those which provide the most valid measure of the component of competence we are testing. The selection of methods will also

need to take into account efficiency as well as efficacy given that we now know that testing time is an important practical issue.

The implications of having to use 4-8 hours of testing time to achieve a reliable assessment of clinical competence are mind boggling to many. Fortunately there are a few interesting new ideas which might relieve some of the strain. For example, if the number of raters is a limiting factor, little reliability is lost by using one rater instead of two whereas much is gained by an increase in the number of stations. Again, if the main purpose of the assessment is to make pass/fail decisions and the majority of the candidates are expected to succeed, a considerable saving in resources could be achieved by sequential testing. Such an approach envisages the use of a short, less reliable test to quite fairly screen out (ie pass) say 70% of the candidates. A different or longer test would then be used to make more accurate decisions on the 30% of students closest to the pass/fail decision point. The same students would eventually pass but less resources would have been required.

Returning to Nijmegen, I am somewhat embarrassed by the naivity of my presentation nine years ago on the assessment of competence. At this time we had very little data on the new approaches to assessing competence which were being advocated as the answer to the reliability and validity problems posed by traditional methods. The major achievement of the last few years, in my view, has been a more critical analysis of what we are doing. It has always been difficult for me to understand why many of my colleagues, who demand such a high standard of the tests they use in their research laboratories or for decision making on their patients, do not insist on the same quality of the tests used to make equally important decisions on their students. Educational tests are often of a low standard but escape criticism.

I therefore, applaud the organisers of this conference for choosing the theme "Scientific Thinking in Medical Education". While the main concern seems to have been with the importance of teaching students how to think scientifically, equally important, in my view, is a need to get teachers to apply the process of scientific thinking to the educational methods we use and particularly to the vital area of student assessment.

REFERENCES


1. **Hart IR, Harden RM & Walton JH** (Eds) (1986). Newer Development in Assessing Clinical Competence. (Proceedings of 1st Ottawa Conference 1985) Heal Publications, Montreal.
2. **Hart IR & Harden RM** (Eds) (1987). Further Developments in Assessment Clinical Competence (Proceedings of 2nd Ottawa Conference 1987) Can-Heal Publications, Montreal.
3. **Teaching and Assessing Clinical Competence** (Proceedings of 3rd Ottawa Conference, 1989) to be published.
4. **Hubbard JP, Levit EJ, Schumacher & Schnabel TG** (1965). An objective evaluation of clinical competence. NEJM, 272,1321-1328.
5. **Neufeld VR & Norman GR** (1985). Assessing Clinical Competence, Springer, New York.
6. **Newble DI, Elmslie RG & Baxter A** (1978). A problem-based criterion-referenced examination of clinical competence. Journal of Medical Education, 53, 720-726.
7. **Newble DI & Swanson DB** (1988). Psychometric characteristics of the objective structured clinical examination. Medical Education, 22, 325-334.
8. **Newble DI, Hoare J & Elmslie RG** (1981). The validity and reliability of a criterion-referenced examination of clinical competence. Medical Education, 15, 46-52.
9. **Newble DI** (1981). The definition of clinical competence. In "Examination in Medical Education" Metz JCM, Moll J & Walton HJ (Eds), 54-57. Wetenschappelijke uitgeverij Bunge, Utrecht.
10. **Newble DI** (1981). The evaluation of clinical competence. In "Examination in Medical Education" Metz JCM, Moll J & Walton HJ (Eds), 77-85. Wetenschappelijke uitgeverij Bunge, Utrecht.
11. **Newble DI** (1988). Eight years experience with a sturctured clinical examination. Medical Education, 22, 200-204.
12. **Norman GR** (1988). Problem solving skills, solving problems and problem-based learning. Medical Education, 22, 279-286.
13. **Van der Vleuten CPM & Swanson DB** (1989). Assessment of cilinical skills with standardized patients: state of the art (to be published).
14. **Wakeford R** (Ed) (1985). Directions in Clinical Assessment (Report of 1st Cambridge Conference) Cambridge.