Validity of final examinations in undergraduate medical training

Cees van der Vleuten

Most medical schools, particularly those in the United Kingdom, have a final examination at the end of undergraduate medical training. Although the format of these examinations has been changed recently by the introduction of newer types of assessment such as objective structured clinical examinations, medical educators are still questioning their validity and worth. I believe that there is a strong case for better continuous assessment during undergraduate training and less reliance on final examinations.

Summary points

- Even with modern forms of assessment, final examinations are of questionable reliability and validity
- They are of limited educational value to students because there is little opportunity for feedback and correction
- The effort spent on running final examinations would be better invested in improved continuous assessment during training
- Continuous assessment through "clinical work samples" is a promising new method of assessing medical students

Department of Educational Development and Research, University of Maastricht, PO Box 616, 6200 MD Maastricht, Netherlands
Cees van der Vleuten
professor of education
c.vandervleuten@educ.unimaas.nl

BMJ 2000;321:1217–9

Functions of the examination

The final examination has at least two functions—an accountability or selective function and an educational one. First and foremost, the final exam should provide a guarantee to society that the training programme delivers competent doctors. It should be able to identify any students who are unfit to practise, so that they can be prevented from doing harm, and to license competent students who are ready for further practice and training. With regard to the educational function, proponents claim that the requirement to sit a comprehensive examination at the end of training means that students revise and recapitulate what they have learned throughout the course, a process which leads to a more integrated understanding of the knowledge and skills they have acquired. Let me examine these functions in greater detail.

The selective function

The outcome of the final examination should predict whether a student will be competent. The process should prevent false negative results and, in particular, false positive ones. In other words, the final examination should be reliable and valid.

Reliability and validity

Reliability refers to the precision of measurement or the reproducibility of the scores obtained with the examination. However, all kinds of "noise" can affect the measurement and therefore the reliability. Several statistical theories have been developed to estimate the reliability of an instrument. The most widely used is classical test theory. This provides reliability coefficients, expressed on a scale of 0 (no reliability at all) to 1 (perfect reliability) as indices of precision. The reliability coefficient may be interpreted as a correlation coefficient between this measurement and a hypothetical remeasurement taken under similar conditions.

Validity refers to the extent to which a measurement actually measures what it is intended to measure. Validity, unlike reliability, cannot be expressed in a single coefficient; it is a conceptual term that takes several forms, as described in the box.

Content specificity

There is overwhelming evidence that the reliability of measurements of clinical competence is hampered by the fact that competence is content specific. Achieving competence in one area (for example, in one clinical case) is not a good predictor of competence in another, even if the areas are closely connected.[1] This may not be surprising when content areas are very different (for example, with cardiology and paediatrics), but the problem also holds within specific content areas. Knowing how a candidate has handled one patient's problem is not a good predictor of how he or she will deal with another, even if it is a related problem. Wide sampling of topics across content areas is therefore imperative. This can be achieved easily with efficient testing formats such as multiple choice questions, but it

Forms of validity

Content validity: When an examination is carefully designed through good selection and weighting of the topics to be assessed it is described as having content validity.

Construct validity: A measurement's ability to differentiate between groups with known differences in ability, such as beginners and experts in a particular area, is often called construct validity. This is because a theoretically predicted outcome of an experiment—in this case the differentiation between groups—underpins the "construct" being measured.

Convergent, divergent, and predictive validity: Validity can also be shown by the strength with which scores for one measurement are related to other measures. When two measures are expected to quantify similar constructs, the correlation between their scores is taken as an index of convergent validity. Similarly, when the measurements should quantify different aspects, the correlation is a reflection of divergent validity. When the measure is used to predict an outcome in the future such as professional success after graduation, the term predictive validity is used.

is difficult to accomplish with testing methods such as computer simulations and objective structured clinical examinations. The clinical examination format of the single long case or just a few short cases is equally inefficient in ascertaining a student's width of knowledge. Furthermore, if another factor, such as the examiner, is seen to influence measurement, a large sample of that factor (that is, more examiners) is needed too.

One point in favour of final examinations is the fact that they are usually comprehensive. Most have several components and generally include written and clinical sections. Sometimes each component has several subcomponents. The fact that a final examination consists of a battery of measurements argues in favour of its reliability since the combination of different components will cover a broad spectrum of competencies. However, the final examination cannot reliably decide a student's competence in relation to an entire curriculum. The final examination is actually a "wrapping up" of the entire learning programme, which means that the sample of competence being tested should be representative of or generalisable to a very wide field of knowledge. If good content sampling in a narrow domain is difficult because of the problem of content specificity, think how much harder it must be in a wide domain.

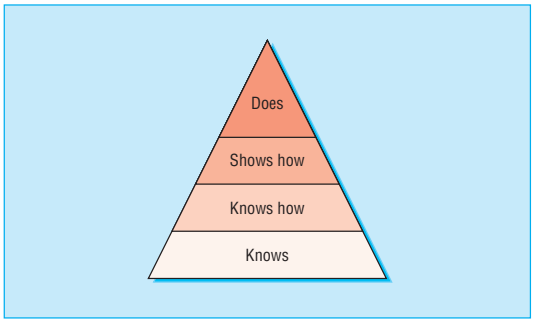
Decision errors

Any discussion on reliability of final examinations remains academic, since there is no good evidence available. Evaluating the reliability of a test that has several components requires special statistical analysis that allows composite estimation, such as that described by Hays.[2] These studies are lacking, and we clearly need more research in order to acquire an empirically verified estimate of the reliability of final examinations. However, it must be remembered that every test has its limitations. Even a dependable test, let us say one with a reliability of 0.80, will lead inevitably to a sizeable number of incorrect decisions. This depends on the test's failure rate: with a failure rate of 5%, the percentage of misclassifications is 5%, but with a rate of 10%, 9% will be misclassified.[3] In all, decision errors in final examinations are inevitable—and probably sizeable.

Measuring clinical competence

The validity issue raises questions about whether we are measuring the appropriate things in the final examination. What medical educators could measure is illustrated by a simple and elegant conceptual model of clinical competence devised by Miller and depicted in the figure.[4] Miller conceives of competence as a pyramid. The base of the pyramid consists of factual knowledge. One level up, Miller describes the ability to use knowledge in a particular context as "knows how." This comes close to clinical reasoning and problem solving. At a higher level, "shows how" reflects the person's ability to act appropriately in a practical situation and describes hands-on behaviour in a simulated or practice situation. The "does" level refers to actual performance in habitual practice. The higher the skills being tested in the pyramid, the more clinically authentic the assessment needs to be. It is arguable that with training, higher levels of the pyramid are reached and that these should be covered in a final examination.

Most final examinations have a written component. Having studied some of these examinations, I would argue that many do not measure more than the knows how level of Miller's pyramid. They are restricted to factual knowledge, even if they are meant to measure application and problem solving. Next to the written component, many examinations have a clinical component that often comprises oral examinations such as short and long cases. If these oral examinations do not include observation of candidates working with a patient, and many do not, they cannot measure any order of competence higher than Miller's knows how level.



Miller's pyramid of clinical competence

Objective structured clinical examinations

The objective structured clinical examination has become very fashionable in recent years and has been included in many final examinations. This assessment concentrates on hands-on clinical behaviour in which examiners are asking the student to show how. However, the objective structured clinical examination, as originally proposed, measures clinical skills in isolation and over short periods (for example, examination of the knee in five minutes or less),[5] which is not a valid representation of clinical reality at the end of an undergraduate curriculum. A clinically authentic problem is one that requires a student to integrate a particular skill with the clinical problem at hand and to take action to manage the patient's problem further. More integral patient problems would approximate the real clinical encounter much better and would enable examiners to ascertain whether the student is operating at the shows how level of Miller's pyramid. Most final examinations seem to require candidates to perform at lower levels of clinical competence, and the candidates' habitual performance in practice (the does level) seems to be the level least well covered in assessment.

Evaluating habitual performance

If examiners wish to assess students at the highest level of Miller's pyramid they need to evaluate the student's habitual performance in everyday practice. A clinical supervisor, who judges a student's general competence at the end of an attachment or clinical rotation, usually undertakes this assessment. However, these ratings are quite unreliable and usually not very informative to the student.[6,7] In fact, it was the lack of reliability of these clinical ratings that triggered the development of the objective structured clinical examination. This examination introduced checklists and the assessment of performance in several situations in order to standardise examination conditions and improve reliability. Some surprising research findings on objective structured clinical examinations have recently shown that these could point to a new way of assessing

habitual performance. There is consistent evidence that a holistic judgment—one made through qualitative judgments measured on a rating scale—is as reliable as a checklist incorporating detailed behavioural items.[8 9] It seems that holistic and actuarial judgments are equally reliable when clinical performance is rated in a concrete situation that is limited in scope and time. However, one requisite, which is true for any measurement, is a wide sample of situations and of examiners.

Sampling students' work

When the situation is less specific in scope and time, such as in a clinical rating covering one judgment over an extended period, the reliability of the holistic judgment is poor. Bearing this in mind, we could assess habitual performance by judging students on a number of occasions while they do their work on a daily basis. The assessment could be simple and holistic, perhaps even generic, judging various dimensions of the clinical encounter such as history taking, differential diagnosis, physical examination, management, and communication, and the assessor would not have to spend much time on the assessment. With sufficient opportunities for assessment, preferably carried out by different assessors so that the examiner sample size is also increased, we could gather inexpensively reliable, valid, and informative data about a student's highest level of competence. These "clinical work samples" would also include an educational component since they would involve more frequent direct observation and provision of feedback. Although it might be assumed that observation and feedback are the core of any apprenticeship learning situation, research has shown that this is typically lacking in clinical rotations.[10 11] However, assessment through clinical work samples would run contrary to the typical model of the final examination in which all the required information has to be assembled at a single, relatively short moment in time.

The educational function

Examinations drive students' learning; this law describes one of the strongest relationships in education. Students wish academic success, academic success is defined by examinations, and students will therefore do anything to maximise their chance of success (usually with the least effort in order to cope with competing interests). The argument that students will repeat and integrate their knowledge and skills as a result of preparing for the final examination is a valid one. However, the strength of this argument depends on the relevance of what they prepare and that will depend on the quality of the examination in relation to its objectives. The difficulties in ensuring that higher levels of the competence pyramid are represented in most final examinations which I described earlier show that the optimum relation between quality and objectives is hard to achieve.

A window of time

A second consideration that puts the argument into perspective is whether the demonstration of competence at a particular moment is really what we are after. The philosophy implicit in the educational argument in favour of final examinations is that of "mastery learning." This is the idea that once a student has shown that he or she is competent at a certain moment in time he or she will stay competent—rather like being "immune for life." Such a belief is questionable, particularly where knowledge and skills are being acquired for an occasion that demands committing a great deal to memory. Information which is memorised for an examination and is not frequently used or repeated thereafter is soon forgotten.[12]

Speed of change

What is learned today is outdated tomorrow. This is a final sobering argument against claiming too great an educational function for the final examination.

Epilogue

In summary, final examinations seem to be changing gradually to include more relevant tasks and skills that are appropriate for the graduating medical student. However, some relevant areas are still not being covered and important aspects of clinical competence assessment are still lacking. In particular, we need to include more professionally authentic assessment to judge higher levels of competence, even where the final examination includes an objective structured clinical examination. Furthermore, it is clear that using the final examination as a selective tool requires caution, and sizeable decision errors should be taken into account. More research is needed in this area. Similarly, the educational argument that learned material is synthesised by the student and learning is stimulated is only true if the final examination really represents the objectives of the curriculum, and I have argued that this is only partly the case. In all, final examinations are unable to perform adequately their selective and educational functions. In the ideal training programme, with a careful and continuous assessment programme throughout, final examinations would be unnecessary. Continuous assessment fosters continuous learning—the two are seamlessly related. Furthermore, the wealth of information gleaned and retained from a continuous and longitudinal assessment programme can never be replaced by a final examination that occurs at a single moment in time.

Competing interests: None declared.

1 Swanson DB, Norcini JJ, Grosso LJ. Assessment of clinical competence: written and computer-based simulations. *Assess Eval Higher Educ* 1987;12:220-46.
2 Hays RB, Fabb WE, van der Vleuten CPM. Reliability of the fellowship examination of the Royal Australian College of General Practitioners. *Teaching Learning Med* 1995;7:43-50.
3 Magnusson D. *Test-theory*. London: Addison-Wesley, 1966.
4 Miller GE. The assessment of clinical skills/competence/performance. *Acad Med* 1990;65:S63-7.
5 Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). *Med Educ* 1979;13: 41-54.
6 Gray JD. Global rating scales in residency education. *Acad Med* 1996;71:S55-63.
7 Streiner C. Clinical ratings—ward rating. In: Shannon S, Norman G, eds. *Evaluation methods: a resource handbook*. Hamilton, ON: Program for Educational Development, McMaster University, 1995:29-32.
8 Regehr G, MacRae H, Reznick R, Szalay D. Comparing the psychometric properties of checklists and global rating scales for assessing performance on an OSCE-format examination. *Acad Med* 1998;73:993-7.
9 Rothman AI, Blackmore D, Dauphinee WD, Reznick R. The use of global ratings in OSCE station scores. *Adv Health Sci Educ* 1997;1:215-9.
10 Jolly BC. Bedside manners: teaching and learning in the hospital setting [dissertation]. Maastricht: University of Maastricht, 1994.
11 Remmen R, Denekens J, Scherpbier AJJA, van der Vleuten CPM, Hermann I, Van Puymbroeck H, et al. Evaluation of clinical skills training during clerkships using student focus groups. *Med Teacher* 1998;20: 428-31.
12 Semb GB, Ellis JA. Knowledge taught in school: what is remembered? *Rev Educ Res* 1996;64:253-86.

(Accepted 18 July 2000)