

Quality Control: Assessment and Examinations

Qualitätskontrolle: Bewertung und Prüfungen

Abstract

Reaching for an optimal quality of medical examinations is important. They serve as a predictor for future performance. Predictions should always be based on accurate assessments of the present situation. Therefore a lot of attention must be directed towards the quality of the examinations used. This quality can be expressed as validity (does the examination really test the competence that one wants to measure?) and reliability (is the score a correct representation of the amount of competence of the candidate?). A major error source concerning these parameters lie in inadequate sampling of the items. Too small samples of items lead to inadequate reliabilities and thus to inadequate validities of the examinations used. The most common reason for this is an overconcern with format of the items rather than content due to the belief that some item or examination formats are intrinsically superior to others. Extensive research shows, however, that the content of the item is essential and not the format. Quality control of assessment lies in a careful and scientific scrutiny of the choice of the examination system, the items that will be put into the examinations, the method to set the cut-off score and the regulations concerning the examinations.

Zusammenfassung

Es ist wichtig eine optimale Qualität der Prüfungen im Medizinstudium anzustreben, da sie als Voraussage über künftige Leistungen dienen. Solche Voraussagen sollten immer auf einer korrekten Bewertung der gegenwärtigen Situation basieren, daher ist der Qualität dieser Prüfungen ein hoher Grad an Aufmerksamkeit zu widmen. Diese Qualität kann als Validität (mißt die Prüfung wirklich die Qualifikationen die wir messen wollen?) und Verlässlichkeit (macht die Prüfungsnote wirklich eine richtige Aussage über die Qualifikation des Prüflings?) ausgedrückt werden.

Eine wichtige Fehlerquelle in Bezug auf diese Größen ist eine zu geringe Zahl von Prüfungsfragen. Zu wenige Fragen bewirkt eine geringe Verlässlichkeit und als Folge eine inadäquate Validität der Prüfung. Dies ist meist die Folge einer allzu intensiven Auseinandersetzung mit der richtigen Form anstatt der richtigen Inhalte einer Prüfung, aus dem Glauben herraus, daß gewisse Prüfungsformen von vornherein anderen überlegen sind. Viele wissenschaftliche Untersuchungen haben aber gezeigt, daß Inhalte der Prüfungsfragen und nicht die Prüfungsform das ausschlaggebende sind. Eine Qualitätskontrolle der Bewertung erfordert eine gründliche wissenschaftliche Überprüfung des gesamten Prüfungssystems, der verwendeten Prüfungsfragen, der Wahl der Kriterien um zwischen Genügend und Nichtgenügend zu differenzieren und der Prüfungsordnung.

1 Introduction

A well accepted fact about assessment of medical competence is the importance of the use of good and appropriate examinations. This not only involves society but also the faculty and the students. The power of examinations to precisely discriminate between the competent and the non-competent students is crucial to all parties involved. All three suffer both from unjustified pass-decisions as well as unjustified fail-decisions.

The competence of the doctors working in the health care system is important to society, not only for the alleviation of human suffering from disease but also to reduce the economical loss due to costs related to illness. The education of students to become doctors, however, must be efficient to optimally reduce education-related costs. This should place a pressure on the faculties to obtain high efficiency rates in their education, whilst ensuring the quality of the graduates. The developments in Europe leading to the vanishing of borders will probably only increase this pressure.

Unjustified fail-decisions can have financial, motivational and social consequences for the students. Unjustified pass-decisions may have negative consequences too. They may not only lead to a loss of faith in the examinations by the students but also to an accumulation of minor study problems to a large one later on which may then prove insurmountable. It is clear that an unjustified graduation may lead to even bigger problems.

Based on all of the above one would expect that a large amount of time and resources are used to optimize the quality of the examinations used. The opposite, though, is often true: in many of the decisions concerning examinations and cut off scores the weight attributed to tradition and intuition is larger than the weight that is attributed to the results of the many studies conducted in this area. This article tries to give a brief summary of the results with their implications for most examination formats in use. First, however, a brief explanation will be given about the framework in which this summary will be placed.

2 Examinations as predictor of future performance

All examinations have the final objective to contribute to a prognosis of future performance. Four important error sources jeopardise this objective. The first one concerns the difference between what physicians are able to do (competence) and what they actually do during their daily practice (performance).¹ The second problem is the fact that knowledge and competence tend to decrease with time, if not reactivated at regular intervals.²⁻⁵ A third problem lies in the question of whether the entity that is being measured may serve as a good predictor of future medical performance. The inaccuracy of scores of many measurements forms the fourth problem. The latter renders it impossible to

make a reproducible pass-fail decision. This is especially the Case with those students scoring close to the cut off score; students that passed could very easily have failed and vice versa.

The first two problems mainly occur due to poor planning of examinations and flaws in examination regulation, the third and the fourth are related more to the actual content of the examination.

3 History of research of assessment

3.1 Focus on medical competence

In many countries (e.g. the US and Canada) multiple choice questions have been widely used to test medical competence. In the 1970s, however, a growing dissatisfaction with this question format started a search for other test formats. The commonly accepted ideas about medical competence were based on a concept that medical competence consisted of a collection of separately measurable entities. Many different entities were suggested: factual knowledge, comprehension, problem-solving ability, clinical reasoning, judgement, decision competence, technical skills, communication skills, attitudes, etc. As many measurement instruments as examination formats were developed.⁵⁻¹⁰ Validity, i.e. congruence in the rankings of student from best to worst, were used as an indication for validity. If these correlations were low the conclusion would be that the every test measured a different trait. Mostly, however, moderate correlations were found. These moderate correlations were still considered to be an indicator that indeed a different trait was being measured.¹¹⁻¹⁴ A confounder, however, is the fact that unreliability of the tests decreases the correlation. So, the moderation of the observed correlations could be attributed to the unreliability of the tests. An estimation must then be made of the correlation when the reliabilities would be ideal. Indeed, studies in which the correlations were corrected for this unreliability show considerably higher correlations.¹⁵⁻¹⁷ It is uncertain what this means. Highly correlating variables do not have to originate from the same entity (body weight and length are highly correlated but are different entities), but on the other hand variables that originate from the same entity do correlate highly (length in centimetres and inches correlate highly). Nevertheless, it appears that the amount of unique information about the competence of students obtained using one examination method (instead of the other) may be relatively small. Especially in examination formats that differ a lot from each other (like performance-based tests and written tests) this seems to be counter-intuitive. It has been suggested also that different cognitive levels are measured by different question formats. Open-ended questions were thought to be superior to multiple choice questions, because an open-ended

question would elicit spontaneous generation of knowledge, while recognition of the correct answer suffices for multiple choice questions. Recognition of the correct answer is called the cueing effect. This cueing effect has been the object of many different studies since the 1960s.²⁻²⁸ Almost all studies yield differences in mean scores but again they show high true correlations. This must thus lead to the same conclusion as drawn previously, it is unsure whether open-ended question tap into another level of cognitive skills or not, but the information obtained by their use is highly similar to that obtained using multiple choice questions.

In summary there are strong indications that the format of the question is of minor importance, and that the content may play a much more important role.

3.2 Domain specificity

Many Case-based examination types have been suggested and studied mainly in the 1970s and 1980s.^{3,7,8} In these examinations the aim was to emulate reality as much as possible. The student had to work his/her way through a Case completely, and all his/her decisions were judged. Besides scoring problems (how much credit should be awarded for what decisions),^{18,19} and construct validity problems (experienced physician scores were equally high or lower than novices),²⁰ a reliability problem existed. It appeared that the ability of a candidate to solve a Case was highly dependent on knowledge about the particular problem, and knowledge is highly dependent on the specific content of the Case.^{21,22} This so-called, 'domain specificity' indicates that the score a student obtains on one Case has an extremely low correlation with the score that would be obtained on any other Case (even within the same subject or domain). Therefore, to make an examination sufficiently reliable, many different Cases must be included. When considering an examination as a sample of all possible questions that could be asked, it is clear that this sample must be large enough. However, the feeling still exists that only a limited number of questions (in written examinations) or only a short period of time (in orals) may suffice in examinations. It is considered more important to elaborate on a specific topic than to ask about as many different topics as possible. In examinations, however, many different items should be asked to obtain an adequate reliability.^{21,22}

In addition to that one must bear in mind that a measurement cannot be valid if it is not reliable, therefore a reliability problem is always a validity problem.

3.3 Influence on study behaviour

Perhaps the most important reason for a teacher to subject their students to examinations might be the fear that students would not work hard enough otherwise. Only little attention is paid to the possible consequences that the ex-

aminations could have on study behaviour. Even in the literature only a few articles can be found reporting systematic studies about this topic.²⁹⁻³¹ A reason for this could be that some of the effects are so obvious that they could not lead to good hypotheses. Also, these studies often are difficult to conduct logistically. Finally, many students perceive differences in their preparation for different examinations, but they do not show in the results.

A conclusion common to the aforementioned studies, however, is that in their study strategies students are strongly led by the content and format of the examination they expect. The practical consequences for the different examinations may be divided into two categories: examination formats (orals, written tests, and performance-based examinations) and planning and regulation of examinations. It is beyond the scope of this article to discuss them all extensively, therefore only some issues will be used as an illustration. The examination formats will be discussed first.

4 Examination formats

4.1 Oral examinations

Many teachers consider this examination type a good measurement tool that enables them to probe to a deeper level of the competence. This would then give a better insight into the knowledge or problem-solving skills.

Reliability of these unstructured oral examinations, however, often is extremely low. Causes for this can be identified easily: subjectivity of the examiner (lenient or harsh), the 'deeper probing' only allows for a small number of Cases to be submitted to the candidate thus increasing the influence of the domain specificity, personal feelings between examiner and candidate, etc. Furthermore, the so-called halo effect may occur: after a short period the examiner has formed an opinion about the candidate and this opinion is not likely to be changed.

As a result of this the validity of unstructured orals cannot be high. The scores will represent all kinds of factors that do not relate to the medical competence of the candidate, but to other sources of errors.

The solution just to test longer and thus increase the sample will probably not work. Firstly the halo effect will interfere, because the opinion formed by the examiner will not be likely to change after some hours. Secondly, in some instances, the prolongation needed will be so enormous (up to several Days of testing time) that it would not be feasible.²² What does work is to add structure to the examination and to question only the essential issues of a Case instead of the whole Case.

Another problem is the fact that in oral examinations too often include factual knowledge. In view of the teacher time required for orals, this is an inap-

propriate method. This time could better be spent on careful production and review of written examination material.

4.2 Written examinations

In written examinations the teachers often prefer open-ended questions. The most often suggested disadvantage of multiple choice questions is that they would be suitable only for testing of trivial factual knowledge. Indeed in the past they have often been used for this. Developments over the last decade have led to a much broader application of multiple choice questions, focusing more on decisions.³³⁻³⁵ Again comparisons between well constructed open-ended and well constructed multiple choice questions show many more similarities than differences.

A further advantage of multiple choice questions is the short response times. This enables more items to be asked per hour of testing time. In view of the domain specificity problem multiple choice tests yield more reliable scores per hour of testing time than open-ended questions.

Multiple choice questions appear more difficult to produce, but the fact that a well constructed open-ended question implies also a clear and explicit answer key is not taken into account. A well written open-ended question therefore probably costs an equal amount of time. The scoring of multiple choice questions, however, is less time consuming, certainly when using computers.

The reliability problems described above are true especially for essay examinations. They constitute only a very small sample and are often corrected only by one examiner. It is practically impossible to give a well-defined pre-fixed answer key. For the testing of medical competence these kinds of tests should therefore be discouraged strongly. If, however, the testing of writing ability is the sole purpose of the examination an indication for this examination type may be present. A prerequisite for this is of course a sufficient literary competence of the corrector himself.

4.3 Performance-based examinations

Domain specificity is a major problem in any examination that tests only a small number of Cases. They then yield unreliable and invalid scores. Unreliability is a two-way street. One does not know whether failing of a student was justified, but one also does not know whether the passing of a student was justified. Structuring of the assignments and measures to broaden the sample can form a solution for this problem. That is the reason why in many institutions the tradition of the examination patient is replaced with objective structured clinical examinations (OSCEs), in which a student must complete a circuit of different rooms with different examiners and different assignments with

different simulated or real patients.⁹ The examiners use structured checklists on which they can tick every item performed by the student. Sampling now occurs of assignments, examiners and patients. Caution must be kept not to overstructure the lists and to run the risk of trivialising them.³⁵

5 Planning of examinations

5.1 Final examinations

A risk of holding final examinations in each subject is the induction of unwanted study strategies. Most people will remember from their own study the 'cramming' for an examination in order to pass it. After the examination the subject matter does not have to be repeated and the student is 'immune' for the rest of the curriculum. This 'immunity principle' is not ideal, since knowledge has a tendency to decrease if it is not reactivated regularly.²⁵ In a somewhat exaggerated style one could say that students 'polish up' for an examination, pass it and subsequently try to forget as much as possible to be able to optimally prepare for the next examination. Final examinations may therefore not induce a continuous study behaviour but more a sort of hurdle run. A system of multiple examinations testing previous learned subject matter in an integrated way, in combination with a rule to combine all the scores would have a more positive effect on the acquisition of knowledge of the students.

5.2 Repeat examinations

Repeat examinations seem fair to the students, but this can be questioned. Since a repeat examination is in fact a repeated measurement of the same entity it suffers from a statistical problem. In every measurement a certain chance of an incorrect decision is included (like in every lab test a 5% chance of either a false-positive or false-negative result is included). Repeating the measurement over and over again may produce a summation of these error chances. The risk of a false-positive result increases therefore with the number of repeat examinations. A second problem is the fact that they cost teacher time, that could better be spent on quality management of the normal examinations. In this sense they harm the other students. A third problem is the induction of a minimalist approach in students. They can allow themselves just to try the first examination to 'explore' the territory because there is always the possibility of a repeat examination.

Competition between examinations (i.e. the fact that when two or more examinations are held in about the same period) students will tend to prepare optimally for the examination that is considered the most important) constitutes a fourth problem. Finally, they can allow for an unwanted study

behaviour. Periods of relative 'idleness' can be compensated for by periods of cramming.

The solution of the combined examination suggested above would avoid most of these problems.

6 Construction of examination material

Writing questions or assignments is a difficult task. Certainly in this area quality assurance is important, because badly formulated questions can easily lead to false-positive and false-negative results. These problems are not confined to any one question format, since open-ended and all kinds of multiple choice questions have their own difficulties. Describing all pitfalls is not the aim of this article, but some remarks can be made.

Writing a good question involves more than the use of normal oral language. The possibility of misinterpretations should be avoided. Questions that are ambiguous can lead to incorrect answers in students who actually do possess the knowledge.

On the other hand, too absolute terminology (like „never“, „always“, „must“ etc.) or too open terminology (like „can“, „is possible“, etc.) can often lead the student to the correct answer. The student may use these leads to eliminate possibilities in multiple choice questions or in true-false questions.

Open-ended questions that do not sufficiently indicate the type of answer expected may also lead to incorrect answers from sufficiently knowledgeable students. But open-ended questions that do not restrict the length of the answer invite the students to write down complete 'stories' in the hope of gaining some points. Open-ended questions without a pre-fixed answer key can lead to arbitrariness in the scoring.

It is therefore advisable to subject all examination material to a review process. This can be done either by reviewing the material, after some delay by the author him or herself, or by a colleague.

7 Conclusion

The construction of examination material is a time-consuming and not always pleasant task. The importance of good assessment and the necessity for optimal efficiency are clear. Unfortunately common beliefs and empirical data sometimes contradict each other. This article has tried to summarize some of the major findings in the literature and their implications for daily practice. Fortunately more research is being carried out in many different places. As communication between teachers of different faculties and in different countries increases, this hopefully will lead to many new insights in the testing of medical competence.

References

- 1 Reihans, J.J., Sturmans, F., Drop, M.J., van der Vleuten, C.P.M., Hobus, P.: Does competence of general practitioners predict their performance? Comparison between examination settings and actual practice. *British Medical Journal* 1991; 303: 1377-80.
- 2 Meskauskas, J.A., Webster, G.D.: The American Board of Internal Medicine Recertification Examination: process and results. *Annual Internal Medicine* 1975; 82:1-5.
- 3 Norcini, J.J., Lipner, R.S., Benson, J.A., Webster, G.D.: An Analysis of the Knowledge Base of Practising Internists as Measured by the 1980 Recertification Examination. *Annual Internal Medicine* 1985; 102:385-389.
- 4 Day, S.C., Norcini, J.J., Webster, G.D., Viner, E.D., Chirico, A.M.: The Effect of Changes in Medical Knowledge on Examination Performance At the Time of Recertification. In: *Proceedings of the 27th annual conference on Research in Medical Education*, Chicago: American Association of Medical Colleges, 1988.
- 5 McGuire, C.H., Solomon, C.M.: *Construction and Use of Written Simulations*. Chicago: The Psychological Corporation, 1976.
- 6 Feltri, G.I., Engel, C.E.: The modified essay question for testing problem-solving skills. *Medical Journal of Australia* 1980; 1:79-80.
- 7 Barrows, H.S., Tamblin, R.M.: The portable patient problem pack (P4). A problem-based learning unit. *Journal of Medical Education* 1977; 52:1002-1004.
- 8 Powles, A.C.P., Wintrup, N., Neufeld, V.R., Wakefield, J.H., Coates, G., Burrows, J.: The triple jump exercise: Further studies of an evaluative technique. In: *Proceedings of the 20th annual conference on research in medical education*, Washington: American Association of Medical Colleges, 1981.
- 9 Harden, R., Gleeson, F.: Assessment of clinical competence using an objective structured clinical examination (OSCE). *Medical Education* 1979; 13:41-54.
- 10 Sullman, P., Ruggill, J., Runala, P., Sabers, D.: Patient instructors as teachers and evaluators. *Journal of Medical Education* 1980; 55:186-193.
- 11 Case, S.M.: A New Examination for the Evaluation of Diagnostic Problem-solving. In: *Proceedings of the 20th annual conference on research in medical education*, Washington: American Association of Medical Colleges, 1981.
- 12 McGuire, C.H., Babbot, D.: Simulation Technique in the Measurement of Problem Solving Skills. *Journal of Educational Measurement* 1967; 4:1-10.
- 13 Skakun, E.N., Taylor, W.C., Wilson, D., Taylor, T., Grace, M., Fincham, S.M.: Preliminary Investigation of Computerized Patient Management Problems in Relation to Other Examinations. *Educational Psychology Measurement* 1979; 39:303-310.
- 14 Wolf, F.A., Cassidy, J., Maxim, B., Davis, W.: A Criterion-referenced Approach to Measuring Medical Problem Solving. *Evaluation of Health Professionals* 1985; 8:223-240.
- 15 Norcini, J.J., Swanson, D.B., Grosso, L.J., Shea, J.A., Webster, G.D.: A Comparison of Knowledge, Synthesis and Clinical Judgement: Multiple-choice Questions in the Assessment of Physician Competence. *Evaluation of Health Professionals* 1984; 7:485-499.
- 16 Norcini, J.J., Swanson, D.B., Grosso, L.J., Webster, D.B.: Reliability, Validity and Efficiency of Multiple Choice Question and Patient Management Problem Item Formats in Assessment of Clinical Competence. *Medical Education* 1985; 19:238-247.
- 17 Norcini, J.J., Meskauskas, J.A., Langdon, L.O., Webster, G.D.: An evaluation of a computer simulation in the assessment of physician competence. *Evaluation & the Health Professions* 1986; 9(3):286-304.
- 18 Norcini, J.J., Swanson, D.B., Grosso, L.F., Webster, G.D.: A Comparison of Several Methods for Scoring Patient Management Problems. In: *Proceedings of the 22nd annual conference on research in medical education*, Washington: American Association of Medical Colleges, 1983.
- 19 Bligh, T.: Written Simulation Scoring: a Comparison of Nine Systems. Thesis: *Urbana-Champaign, IL*, University of Illinois.
- 20 Marshall, J.: Assessment of Problem-solving ability. *Medical Education* 1977; 11:3-29-334.
- 21 Elstein, A.S., Shulman, L.S., Sprafka, S.A.: *An Analysis of Clinical Reasoning*. Cambridge MA: Harvard University Press, 1978.
- 22 Swanson, D.B., Norcini, J.J., Grosso, L.J.: Assessment of clinical competence: written and computer-based simulations. *Assessment and Evaluation in Higher Education* 1987; 12(3):220-246.
- 23 Hettiaratchi, E.S.G.: A comparison of student performance in two parallel physiology tests in multiple choice and short answer forms. *Medical Education* 1978; 12: 290 - 296.
- 24 Hurlburt, D.: The relative value of recall and recognition techniques for measuring precise knowledge of word meaning. *Journal of Educational Research* 1954; 47(8): 561-576.
- 25 Maatsch, J.L., Huang, R.H.: An evaluation of the construct validity of four alternative theories of clinical competence. In: *Proceedings of the twenty-fifth Annual Conference on Research in Medical Education*, Washington: American Association of Medical Colleges, 1986.
- 26 Newble, D.I., Baxter, A., Elmslie, R.G.: A comparison of multiple-choice tests and free-response tests in examinations of clinical competence. *Medical Education* 1979; 13:263 - 268.
- 27 Norman, G.R., Smith, E.K.M., Powles, A.C., Rooney, P.J., Henry, N.L., Dodd, P.E.: Factors underlying performance on written tests of knowledge. *Medical Education* 1987; 21:297 - 304.
- 28 Schuwirth, L.W.T., van der Vleuten, C.P.M., Donkers, H.H.L.M.: Open-ended questions versus Multiple Choice Questions, An Analysis of Cuing Effects. In: R.M. Harden, I.R. Hart, H. Mulholland (Eds.): *Approaches to the Assessment of Clinical Competence*, part 2. Norwich: Page Brothers, U.K., 1992.
- 29 Newble, D.I., Jaeger, K.: The effect of assessments and examinations on the learning of medical students. *Medical Education* 1983; 17:165-171.
- 30 Hakstian, A.R.: The Effects of type of Examination Anticipated on Test Preparation and Performance. *The Journal of Educational research* 1971; 64(7): 319-324.
- 31 Stalenhoef-Halling, B.F., van der Vleuten, C.P.M., Jaspers, T.A.M., Fiole, J.F.B.M.: