2011; 33: 224–233 MEDICAL TEACHER

Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference

LAMBERT SCHUWIRTH[1], JERRY COLLIVER[2], LARRY GRUPPEN[3], CLARENCE KREITER[4], STEWART MENNIN[5], HIROTAKA ONISHI[6], LOUIS PANGARO[7], CHARLOTTE RINGSTED[8], DAVID SWANSON[9], CEES VAN DER VLEUTEN[1] & MICHAELA WAGNER-MENGHIN[10]

[1]Maastricht University, The Netherlands, [2]Southern Illinois University School of Medicine, USA, [3]University of Michigan Medical School, USA, [4]University of Iowa Carver College of Medicine, USA, [5]University of New Mexico School of Medicine, USA/Brasil, [6]University of Tokyo, Japan, [7]Uniformed Services University, USA, [8]Rigshospitalet, Denmark, [9]National Board of Medical Examiners, USA, [10]Medical University of Vienna, Austria

Abstract

Medical education research in general is a young scientific discipline which is still finding its own position in the scientific range. It is rooted in both the biomedical sciences and the social sciences, each with their own scientific language. A more unique feature of medical education (and assessment) research is that it has to be both locally and internationally relevant. This is not always easy and sometimes leads to purely ideographic descriptions of an assessment procedure with insufficient general lessons or generalised scientific knowledge being generated or vice versa. For medical educational research, a plethora of methodologies is available to cater to many different research questions. This article contains consensus positions and suggestions on various elements of medical education (assessment) research. Overarching is the position that without a good theoretical underpinning and good knowledge of the existing literature, good research and sound conclusions are impossible to produce, and that there is no inherently superior methodology, but that the best methodology is the one most suited to answer the research question unambiguously. Although the positions should not be perceived as dogmas, they should be taken as very serious recommendations. Topics covered are: types of research, theoretical frameworks, designs and methodologies, instrument properties or psychometrics, costs/acceptability, ethics, infrastructure and support.

Preliminary statements

Not so long ago, the main focus of assessment was on measuring the outcomes of the learning process, i.e. to determine whether the students had acquired sufficient knowledge, skills, competencies, etc. This approach is often referred to as assessment *of* learning. Currently, a second notion has gained ground, namely assessment *for* learning (Shepard 2009). In this view, assessment is seen as an essential and integral part of the learning process. The purpose of this article is not to elaborate further on these developments, nor to take a stance on it. The sole reason for highlighting it is that it makes the distinction between educational research and assessment research less clear. Therefore, it is inevitable that this article contains descriptions, positions and consensus that do not pertain *exclusively* to assessment research but may have bearing on more general educational research as well. So, although the remit for the theme group has been to focus on assessment, it cannot be avoided that some of its content is of more general pertinence.

The terms assessment and evaluation are used interchangeably in the literature, yet can refer to different inquiries. Some languages have one word for both terms making translation difficult. The theme group on Research in Assessment has agreed that for purposes of clarification and consistency, the term assessment will be used to refer to the systematic determination of student/learner achievement and performance. The term evaluation will be used with reference to issues and questions related to programmes, projects and curriculum within which questions and issues of assessment of learners are nested and co-embedded with educational issues and questions related to resources, faculty, general institutional and programmatic outcomes as well as explanations of educational intervention.

Introduction

Medical education as a scientific discipline is still young. Although the two disciplines on which it is founded, educational psychology and clinical medicine, have a much longer scientific history, medical educational research itself did not start as an independent stream before the 1960s. It is now a rapidly changing field seeking its own scientific identity, not in the least because the scientific languages and mores of medicine – as a biomedical science – and educational psychology – as a social science – differ considerably.

Correspondence: L. Schuwirth, Educational Development and Research, Maastricht University, PO Box 616, Maastricht, 6200 MD, The Netherlands. Tel: 31 43 3885731; fax: 31 43 3885779; email: l.schuwirth@maastrichtuniversity.nl

ISSN 0142–159X print/ISSN 1466–187X online/11/030224–10 © 2011 Informa UK Ltd.
DOI: 10.3109/0142159X.2011.551558

Medical education is therefore faced with the challenge of defining itself relationally so that it fits identifiably within the larger context of health professions education and the biological and social sciences because of its unique characteristics rather than in spite of them. It adds a new flavour to the culinary possibilities of professional education.

With this consensus article, we aim to support this development by describing the most important aspects of medical educational research in general, and research into assessment in particular, by offering concrete positions about it and recommendations. We hope to make the reader understand where we are at the moment and what is needed for research in assessment to evolve further. We also seek to explain to readers of various backgrounds (medical and psychological) why medical assessment research is neither completely analogous to biomedical nor to psychological research, but is emerging as a discipline of its own.

This consensus article is intended for those who have an interest in assessment research, either as a 'customer' or as a 'producer'. With the former, we mean anybody who is involved in designing assessments and seeks to support his/her decisions with the assessment literature. With the latter, we mean anyone who is involved in active research.

As said, in this article, we will offer consensus standpoints on issues concerning medical assessment research, and where possible, we will provide suggestions. Standpoints may be phrased in two ways. They may be phrased as goals that are desirable and should be attained as well as possible (as far as the context of the study allows) and standpoints on essential features that must be adhered to in assessment research.

Definitions

In this article, we will use several terms for which definitions in the literature vary. To eliminate ambiguity as much as possible, we will provide definitions of these terms here. This does not necessarily mean that our definitions are exhaustive, nor that they are intended to replace some of the definitions used in the literature.

Theory

When we use the word 'theory', we refer to rational assumptions about the nature of phenomena, based on observations, and subject to scientific studies aiming to verify or falsify the theory. A theory is not necessarily practically useful but it must be useful for understanding of phenomena.

An example of a theory is classical test theory. In this theory, the backbone is the notion that an observed score is the sum of true score (the score a candidate would have obtained if s/he had answered all the possible relevant items of a certain domain) and the error score.

Theoretical framework

With a theoretical framework, we imply a set of related theories that together serve to explain a complicated aspect (in this case concerning assessment). An example of a theoretical framework is the approaches to validity. Validity, for example, has been defined as the minimalisation of construct under-representation and construct-irrelevant variance by Messick (1994) and as an argument-based rationale by Kane (2006). Both views can be seen as theoretical frameworks.

Note that we avoid using the term 'paradigm' here. 'Paradigm' has a distinct meaning with important implications depending on the specific philosophy of science stream in which it is used (compare for example, the views of Kuhn and his successor Imre Lakatos). Therefore, we think that the word should be used with care.

Conceptual framework

Where a theoretical framework tries to formulate a series of related theories to explain optimally a complicated phenomenon, a conceptual framework helps to interpret findings and give directions. An example of a conceptual framework is assessment *for* learning as opposed to its conceptual counterpart assessment *of* learning. In the context of a framework of assessment *of* learning, case-to-case variance can be seen as error, whereas in the conceptual framework of assessment *for* learning, the same type of variance can be seen as meaningful variance (e.g. because it provides the teacher with an entry to stimulate the learning of a specific student, through the identification of strengths and weaknesses).

Types of research

The ideographic description or 'case report'

Often, scientific research is associated almost exclusively with experimental research; the investigator (preferably in a white lab coat) conducts a carefully planned and controlled experiment. Certainly, medical education scientific research is much broader than this. Basically, one could state that essential in scientific research is a planned and structured collection or management of data with the intent to generate or add to, generalisable knowledge (Miser 2005). This bears in it the notion of general relevancy and applicability, often epitomised in the two questions: 'who cares?' and 'so what?' (Bligh 2003).

Despite the above definition of scientific research, an abundant feature in the literature has been the ideographic[1] description of assessment methods and approaches. One of the earliest examples of this – in general education – was given by McCall (1920) who introduced the true-false examination format. Such descriptions can be seen as analogous to the case reports in the medical literature. They may not be in line with the definition of 'a planned and structured collection or management of data' mentioned above but they certainly serve a purpose.

Educational 'case reports' describe innovations without providing much of supportive data; they may describe new instructional methods or assessment tools. For example, one can regard the first publication on the objective structured clinical examination (OSCE) as such a presentation paper or case report (Harden & Gleeson 1979). Yet, it has had an enormous impact on practice, as OSCEs are probably the most widely used assessment approaches for the assessment of

(clinical) skills. More importantly, however, it has given rise to a plethora of studies, which have significantly increased our understanding of the influence of different sources of variance on scores, especially demonstrating the relationship between inter-observer unreliability and inter-case unreliability (Swanson 1987). If we had not had this line of research, it would have been very difficult to support current approaches such as mini-CEX (Norcini et al. 1995). Therefore, educational descriptions of local innovations can be valuable under certain provisos. The most important of these provisos is that the authors describe the implications of their instrument/method/approach. In other words, what uses does it have, how should it be used, which of its elements make it so useful and how should it be used in other contexts? Also, whenever possible, they should provide directions for further research. So, although these types of publications are in origin ideographic, their nomothetic[2] aspects should be optimised.

Recommendation 1: The educational 'case report' should always surpass the idiographic description of an instrument, method or approach and lead to generalisable knowledge. A case report can only optimally contribute to the existing literature if it is supported by a description of the theoretical rationale for the instrument and a discussion hypothesising which generic aspects of it can be used in other contexts or situations.

A 'modern' example of such a study is the paper introducing the multiple mini interview by Eva et al. (2004). This presentation of a new approach is firmly rooted in the literature on admission and selection procedures and the OSCE literature, and builds a bridge between both areas in the literature. It provides the rationale, first descriptive statistics and psychometrics and implications for its use.

Developmental or design-based research

There is always a need to develop new instruments or new approaches to assessment. A new development, however, is more than merely a good idea. The theoretical body and the amount of literature in medical education and related disciplines such as (cognitive) psychology, general education and psychometrics should be ample enough to base an idea on and we take the stance that a good review of the literature and subsequent underpinning of the idea is a condition (or, *sine qua non*) for the development of new ideas (even if this literature serves to support the premise that there is a need for a novel approach). This being said, design-based research is more than just a good idea. It is a series of studies (often qualitative and quantitative) aimed at both expanding on the theory and developing and improving new instruments, methods and approaches. It is both systematic in that it employs rigorous methodologies and flexible in that subsequent studies are tailored to the outcomes of previous ones all in one big iterative process of theory building, design, development, implementation and analysis.

Recommendation 2: Developmental or design-based research should be realised through more than one single study, and be planned as a train of studies building the bridges between the idea, the pilot experiments, the improvements, the use in real life, etc.

Justification research

A type of research that has been abundant in medical educational research is justification research. The epitome of this are the studies aimed at proving that some educational approaches are better than others, or in clinical research proving that a new drug is better than a placebo. In assessment, the most well-known examples are the studies aiming at determining whether open-ended questions are superior to multiple-choice questions. Such research is important, especially to convince many stakeholders about an assessment approach, but it also has its limitations. Justification research is, for example, not strong at providing insight into the underlying processes; the results do not tell us why an approach works or does not work. A second limitation is that the 'justification' may or may not apply to other schools and settings. A clear description of the theory underlining the study design allows others to interpret for their own situation.

Recommendation 3: We take the stance that it is imperative that any justification research is done from one or more certain well-founded and well-described theoretical frameworks. Without theory, the results are often of limited use.

A simple example may clarify this. Many studies in medical assessment have tried to determine which question format is better, open-ended or multiple choice. Many studies have simply correlated scores on a multiple-choice test to those on an open-ended test on the same topic. Typically moderate correlations 0.4–0.5 were found (Norman et al. 1996; Schuwirth et al. 1996). These, however, are uninterpretable results as it is still unclear whether the glass is half full or half empty. Had the research been done from the framework of validity and cognitive psychology, the question would for example have been whether the format determines the type of memory pathways the question elicits more than the content of the question does. Such comparisons show that when the content is similar and the format different correlations are extremely high, and when the content differs and the format is the same correlations are extremely low (Norman et al. 1985). Looking into this further through a think aloud protocol study, using the theory on expertise and its development then shows that the stimulus type (case-based or plain factual knowledge) directly influences the quality of the thinking processes (Schuwirth et al. 2001).

Unfortunately, this limitation is often overlooked, because in medicine, the randomised controlled trial is frequently seen as the best or ultimate scientific approach. Yet in medical education, justification research can only serve to form one link in the chain connecting theoretical scientific findings with practice. Much like studies proving the superiority of one cancer drug over a placebo do not help to gain insight into the fundamental mechanisms of cancer.

In addition, one single big justification study cannot answer complicated questions. Therefore, the often heard question: 'this is all very nice but does this produce better doctors?' is unanswerable with one single study (Translated to the clinical

context it would be: 'This CT-scan and MRI stuff is all very nice but does it improve the health of the national population?). There are good discussions in the medical educational literature of this debate (Torgerson 2002; Norman 2003; Regehr 2010).

Recommendation 4: We take the position that that justification research alone is not able to answer the 'big questions' and needs to be incorporated in a programme of research.

Typical research questions in justification research are: is assessment approach a more valid than B, does assessment approach A lead to better learning than B, is assessment approach A more feasible than B, etc.

Fundamental theoretical or clarification research

Without fundamental theoretical research to clarify the mechanisms of learning and assessment, the actions justified by these studies would be poorly understood, and there would be no theory to build on. There is no typical example of an approach here. Fundamental theoretical research can be qualitative or quantitative but seeks to understand how things work, or why things work (Miscellaneous authors 2001).

It is obvious that for a good research project, be it in medical education in general or in assessment in particular, it is essential to review the existing literature well. Rarely, 'new ideas' have not been discovered or described before. Not reviewing the literature, of course, can lead to unnecessary duplication, provided it ever gets published. This is not to say that a replication study cannot provide the marginal benefit of an additional example of something already demonstrated, but the n^{th} replication may not be the best use of resources (and journal space). Also, good knowledge of the existing literature on the topic may help to sharpen the research question, and focus it better on what is still not known instead of only replicating what is known.

Recommendation 5: Replication studies must/should be prepared by a literature review which identifies unique features and purposes of the study.

Literature descriptions in a research paper are often limited to the existing literature within the field of the investigator, and within the few journals of the specialty. We would urge any researcher to also scan adjacent or comparative fields. Research on assessment in the recent decades has profited much of the research in cognitive psychology on expertise. Current research into workplace-based assessment has more to offer if the research in the business literature is included as well. In this way, the researcher is challenged to be more precise about what his/her study adds to the existing literature; i.e. whether it is something completely new, whether it is a replication of findings in a totally different field, or whether it is a replication study in different context.

Recommendation 6: When designing a fundamental theoretical study, the researcher should not only scan the existing medical education literature but also relevant adjacent scientific disciplines (e.g., cognitive psychology, business literature on appraisal, etc.)

Recommendation 7: We take the position that currently there is a need for more clarification research in assessment. Much of the descriptions so far have been idiographic and have not contributed to the emergence of solid underlying scientific theories. Theory formation in science is essential, because without unifying or at least supporting theories, individual studies cannot be linked together in a meaningful way, results cannot be interpreted meaningfully enough and new studies cannot be planned with sufficient focus.

Theoretical frameworks/context

As stated above, research in educational assessment is rooted in social scientific research to a large extent. In social scientific research, even more than in biomedical research, a clear choice of a specific theoretical framework is needed. As many of the aims of the study are based on theoretical constructs, results cannot be interpreted without a clearly described theoretical framework. There are numerous examples for this. One currently interesting field in the assessment is, for example, the use of human judgement in assessment, especially in workplace-based assessment. Research questions can be approached from the theories of naturalistic decision making (Klein 2008), cognitive load theory (Van Merrienboer & Sweller 2005), or theories on the actuarial value of human judgement (Dawes et al. 1989).

There are several reasons why such theoretical frameworks are desirable. First, they help to focus the research questions, and underpin the operational definitions of the variables or constructs explored. They are useful in helping us understand the implications of the results and conclusion. They help us to clearly stipulate hypotheses than can be falsified or corroborated. Most importantly, however, they serve to link various studies together to a coherent overarching theory or paradigm, either by using studies founded in the same theoretical framework or different studies comparing or juxtapositioning different frameworks. This helps medical assessment research to evolve into a coherent scientific domain in which studies can be planned to form a programme of research and to prevent it from being mainly a domain with anecdotal individual studies.

Recommendation 8: Studies in the field of assessment should whenever possible be conducted from a clearly defined theoretical framework. This framework must be reported in the introduction, be used in the description of the methods and in the discussion.

Recommendation 9: If it is not possible to use a theoretical framework at least a thorough review of the existing literature must have been performed to clearly determine where the specific research is positioned in the existing literature.

Recommendations 8 and 9 may not always be easy to adhere to when designing a study or writing a paper. A suggestion to aid in this is – as an exercise – to try to write the introduction of the study without mentioning the local context in which it was performed and still be able to demonstrate the importance and relevancy of the study (the 'who-cares-and-so-what' question).

If it is impossible to do this, the onus is on the researcher to explain what makes the context of the study relevant to its outcomes, for example, by explaining what it has in common with other institutions or what is different, or what is known in the literature. What it is about your setting what makes it interesting to other ones.

Another very important issue is the theoretical and practical contexts in which the study was performed. Two mainstream contexts at the moment are assessment *of* learning versus assessment *for* learning. The former is aimed at establishing accurately enough whether the student's learning activities have made him/her sufficiently competent. The latter includes the inextricable relationship between assessment and learning. The former mainly approaches assessment as a – psychometric – measurement problem, the latter as an educational design problem.

Study design, choices of methods

Many different methodologies can be chosen to conduct research into assessment of medical competence. Contrary to some beliefs, we would take the stance there is no such thing as one single inherently better methodology. The best methodology is the one that is most suited to answer the research question. It is important that the researcher is able to provide a coherent and defensible rationale as to why the particular methodology was chosen, weighing the advantages and disadvantages carefully (typically in the discussion).

Recommendation 10: Avoid thinking in terms of innate superiority of one methodology over another, but rather as the best methodology is the one that is optimally able to answer the research question.

This is not always easy to achieve. Most of us are brought up within a certain research tradition with its own language and idioms. A suggestion we want to offer is to think through the chosen methodology, to imagine the possible outcomes of the study and then consider critically which conclusion you could draw from them. If the results are inconclusive or there are too many possible competing explanations for the result (confounding for example), try another methodology.

Recommendation 11: As educational research is not easy to conduct, it is always wise to include in your team someone with expertise in the methodology you want to use, rather than to simply assume that anyone with a sound mind can do this type of research.

Instrument characteristics: Validity and generalisability

Research in assessment often involves the use of assessment instruments. For the correctness of the outcomes of any scientific study, the use of carefully designed instruments is a necessary condition. In social scientific research, the instrumentation is not always as standard as in other types of research. We do not have experimental animal models, standard ultra centrifuges, etcetera, but we often have to design our instruments ourselves or have to adapt them from others. This makes it essential that instrument development and description are conducted with the utmost care. Just a collection of questions, for example, do not make a good questionnaire, and just some items do not make a good test.

Two elements are central in the determination of the value of the instrument: validity and reliability.

Validity

In the past century, various approaches to validity of assessment instruments have been used. Central in this discussion, however, is that we aim to assess an aspect (construct) that is not directly visible, but that is assumed to exist and has theorised characteristics. Therefore, a validation procedure for an assessment instrument is always a series of studies evaluating the extent to which the instrument scores help to assess the construct. A validation procedure is– much like a scientific theory – never finished. Instead, it needs to consist of a series of critical studies to determine whether the test actually assesses the construct it purports to measure.

The most obvious – and historically first – major notion of validity is one of criterion validity, i.e. does the assessment result predict performance on a criterion measure well enough. Where possible, this is a convenient approach to validity because it is quite practical in convincing stakeholders who are less well versed in education. But, because we aim at assessing an invisible/intangible aspect, a tautological problem may arise, in that the criterion may be needing validation as well. If we are validating an assessment instrument which is supposed to predict whether someone will be a good professional, we need some criterion to measure 'good professionalism'. This criterion is also a construct and may want validation as well. This would then invariably lead to a sort of Russian doll problem of needing another criterion to validate the criterion, etc. *ad infinitum*. This is the reason why criterion validity as the dominant approach has lost ground.

A second intuitive approach to validity is content validity. In content validity, expert judges carefully evaluate the content of the test and determine to what extent this content is representative of the construct of interest. Although it is inherently an obvious approach and it is easy to explain and defend, the sole use of human judgements is its bottle neck as well. If the judges in the content validation process are the test developers themselves, they cannot be expected to be neutral, but also with independent judges, many possible sources of bias (see for an easily accessible overview (Plous 1993) may exist or the specific choice of the judges may influence the outcome of the process, much equivalent to the problem with Angoff panels (Verhoeven et al. 2002).

The currently dominant notion of construct validation is analogous to the concept of the empirical approach with theory generation, data collection, analysis and refinement or change of the theory. In this view, validation is a process of first explicating carefully the construct one tries to assess, and then collect data to see whether the assumed characteristics of the construct are sufficiently captured by the assessment instrument. An assessment instrument can therefore not be valid in itself; it is always valid FOR a certain purpose. Although this is currently a highly popular view, it suffers from

the same central concern as does the inductivistic approach to science, namely that one never knows whether there are sufficient observations to verify the validity, or put more precisely, whether there is not a valid observation possible that would successfully 'falsify' the validity.

Current views on validity, therefore, are comparable to modern ideas in the philosophy of science, namely that validity must be seen as an argumentation process built on several inferences and theoretical notions. Kane (2006) describes validity therefore as a set of inferences and their strengths. First, there is the inference from observation to score, i.e. how the observation of actions of students are converted to a scorable variable. The second inference is one from observed score to universe score. This is highly equivalent to but not the same as reliability in the standard meaning in the literature. Before one can make an inference about the generalisation from observed score to universe score, one must make theoretical assumptions about the nature of the universe. Internal consistency measures (alpha, split-half reliability, KR, test retest) are all useful approaches to the second inference providing the universe from which the sample was drawn is internally consistent. In such a situation, for example, case specificity is an error source. If, however, the theoretical notion of the construct is one of heterogeneity, internal consistency of the sample is an indication of construct under-representation and therefore of poor generalisability to the universe score. In this situation, case specificity is, instead, innate to the construct. If one were to take the blood pressure of a group of patients during the day and find clear difference between the patients, but no variation between measurements within patients, it would be highly internally consistent, but would not be considered a generalisable sample, simply because the construct of blood pressure is assumed to vary with the moment of the day or previous activities.

The third and fourth inferences are from universe scores to target domain and construct. In other words, is the generalisable score on this test representative for the construct or does it very generalisably measure only one element of the construct.

Messick (1994) has highlighted the consequences of the assessment procedure as an element to include in our thinking about validity. This is an important notion because assessment never takes places in a vacuum and can never be seen disentangled from its (educational) consequences.

Validity is thus never a static but always a dynamic process of collecting data, analysing and refining the instrument to match better the construct. If, in research, an instrument is used that is validated elsewhere in a different context, the research project needs to include sound rationales or data to support that the instrument is also valid for the chosen purpose in the current situation.

Recommendation 12: Instruments are never completely validated; validity is always a matter of collecting evidence to support the arguments for validity. Arguments may be deductive, inductive or defeasible.

Recommendation 13: An instrument is never valid per se. An instrument is always valid FOR something. If necessary, even an instrument validated in another context needs to be validated again for the context in which the specific study was done.

Reliability

Though generalisation of the observed score to the universe score is part of the validity process, the concept of reliability is often treated separately in many studies. Therefore, we want to spend a specific section of this article on reliability.

Assessing individual differences regarding competencies, knowledge, skills and attitudes requires assessment instruments that are capable of capturing these differences and translate the empirically observable differences in the domain of interest into meaningful numbers. Amongst others, the most basic requirements for any measure used in educational context are validity and reliability. Validity, as stated in the previous paragraphs, refers to the degree to which evidence and theory support the indented interpretation of the test scores, Reliability refers to the consistency of the measurement, when it is repeated. Psychometric analysis offers statistics, that can be used to contribute to validity evidence, but also offers statistics to quantify the consistency of an instrument, when it is repeated. Basically three types of inferences can be required:

(1) Would the student obtain the same score on the parallel test as s/he did on the actual test?
(2) Would the student take the same place in the rank ordering from best to worst performing student on the parallel test as s/he did on the actual test?
(3) Would the student obtain the same pass–fail decision as s/he did on the actual test?

Knowing reliability of a measure is important, as reliability influences, for example, the extent to which the different measures can correlate (for instance, to estimate the disattentuated correlations) which is relevant in research settings. Reliability coefficients are also used to calculate a confidence interval around a score, thus determining the score range, taking reliability into account.

Three theoretical approaches to reliability are currently popular, classical test theory, generalisability theory and probabilistic theories (item response theory (IRT) and Rasch modelling).

Classical test theory. The basic principle of classical test theory is the notion that the observed score reflects the results of a true score (the score a candidates deserves based on his/her competence) plus error. This is expressed as the association between the observed score and the score on a so-called parallel test (an equally difficult test on the same topics).

The most popular statistic within CTT is Cronbach's alpha, expressing the consistency of the items used, assuming each item being a 'parallel-test'. This, however, is in the majority of the cases not the most suitable approach (Cronbach & Shavelson 2004). Cronbach's alpha is basically based on the notion of a test–retest correlation and is therefore only a valid approach to express the degree of replicability of the rank order of candidates' scores. As such, it is always an overestimation of the reliability if a criterion-referenced (absolute

norm) approach is used. In this case, the specific mean difficulty of the test is an error source which Cronbach's alpha does not take into account.

Recommendation 14: Cronbach's alpha must not be used to estimate the reliability of criterion-referenced tests, in such cases domain-referenced indices must be used.

Although the literature provides some rules of thumb for the interpretation of alpha (generally 0.80 as a minimum for high-stakes testing), it is more advisable to use reliability to calculate the standard error of the mean and from this a 95% confidence interval around the cut-off score to determine for which students a pass–fail decision is too uncertain. That way, the reliability is compared to the actual data and the robustness of the pass–fail decisions is established. Based on the score distribution and the cut-off score, there are situations in which an alpha of 0.60 gives more reliable pass–fail decisions than in other situations (with other distributions and other cut-off scores) with an alpha of 0.80.

Recommendation 15: Reliabilities should always be interpreted in the light of their influence on the actual data. For example if reliabilities are reported with respect to a summative test, they should always be interpreted in the light of possible pass/fail misclassifications.

Generalisability theory. Generalisability theory expands the approach of classical test theory and is a more flexible theory in that it allows the user to dissect and estimate error variance from various sources. Under the assumption that differences in examinees' scores are partly based on differences in assessed competence (true variance) and partly the result of unwanted sources (error variance), generalisability theory enables the calculation of reliability as the ratio of true score variance to total score variance.

Generalisability theory has additional flexibility in that it allows researchers to specifically include or exclude sources of error variance in the equation. It caters, for example, to both criterion and norm-referenced scoring frameworks. In the former, for example, systematic variance related to certain facets of measurement (e.g. systematic item variance) are included in the equation and in the latter they are left out of the equation. But this flexibility comes at a price. Researchers must be very careful in thinking about the designs they use, which sources of variance to include and which are not to be included, which to treat as random and which as fixed factors, etc. Also, it requires that researchers completely describe the chosen design in any publication and report complete variance component G study tables, because without this sort of complete reporting, results cannot be interpreted or evaluated by the reader and the results cannot be incorporated into meta-analytic synthesis.

Analogous to the procedure mentioned under classical test theory, generalisability theory also enables the calculation of the reproducibility of pass–fail decisions using a so called D-cut analysis. In simple high-stakes summative competency assessments, this can be a better approach to estimating reliability.

A final feature of generalisability theory is the possibility of a decision study (D-study). With such a study, the generalisability can estimated given any number of items, judges, occasions, etc. This is important because it enables us to make implementation decisions (hence the name decision study).

Recommendation 16: When applying Generalisability theory to an instrument, the researchers must have clear conceptions about the nature of the sources of variance; they include and the design used in the analysis.

Recommendation 17: The description of generalisability analyses in any report of the study must be such that an independent research can replicate the study. So a comprehensive description of the sources of variance, treatments of fixed and random factors, designs and complete variance component tables must be provided.

Probabilistic theories (IRT and Rasch modelling). A disadvantage of both previous models is that they can only make estimates which are dependent of the particular group of students and that they are therefore not able to estimate the difficulty of a test or of the individual items independently. IRT is a theory that is able to make these estimates independently of the group of test takers. There are basically three models that can be used. The first and simplest is the so-called one-parameter model. In this model, per item, the relationship is determine with the probability of a correct answer and the ability of the candidate (Figure 1).

But since difficulty alone is not enough as parameter to select and manipulate tests, a two-parameter model includes discriminatory power as well. In this model, not only the probability of a correct answer, given the test taker's ability, is included but also the power of the item to discriminate between two test takers of different ability levels (Figure 2).

If three-parameter models are used, the offset is included. This is not precisely the same as a random guessing chance but is similar (Figure 3).

It may be obvious that the more parameters are included in the model, the more pre-test data are needed. As a rule of thumb, 200 test takers can be enough to pre-test a one-parameter model whereas up to 1000 would be needed for a stable fit of a three-parameter model.

Although IRT modelling is a strong and very flexible theory, it requires extensive pretesting and the underlying statistics are complicated and need to be understood well enough to prevent incorrect use and false conclusions drawn on the test.

Recommendation 18: Do not use IRT modelling unless you are sure your data fit the assumptions and you have the necessary expertise in your team to handle it.

A final issue in reliability is the misconceptions that exist about the relationship between objectivity/subjectivity on the one hand and reliability/unreliability on the other. It is often assumed that subjective parts in assessment are automatically unreliable. This, however, is not the case. Reliability or generalisability is a matter of sampling. Too small or too one-sided samples may collect so-called objective information but may still be unreliable. For example, a 1-item

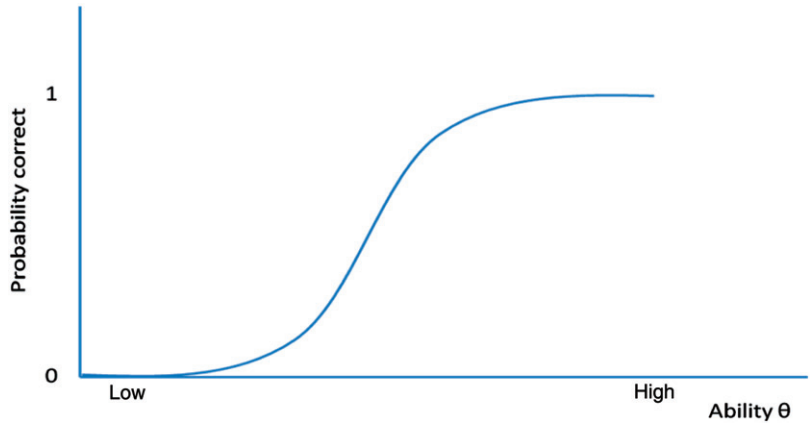

Figure 1. An example of a one-parameter model relationship between the probability of a correct answer and the ability of the candidate.

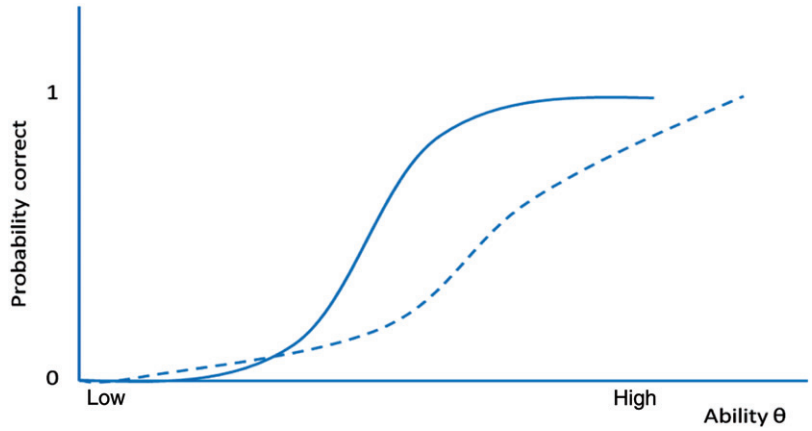

Figure 2. An example of a two-parameter model relationship between the probability of a correct answer and the ability of the candidate.

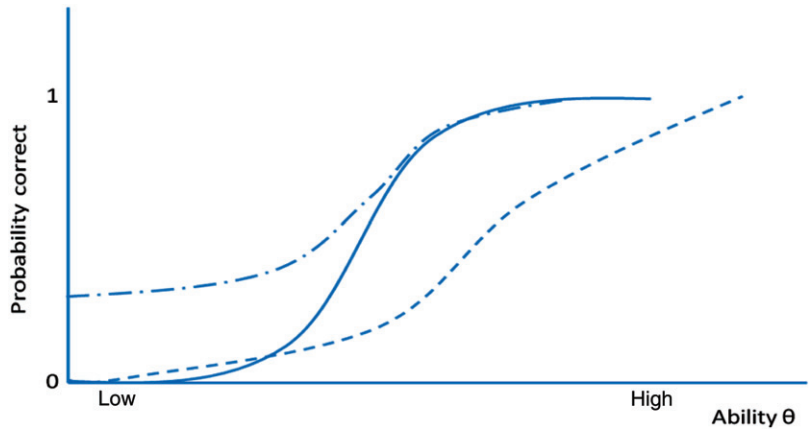

Figure 3. An example of a three-parameter model relationship between the probability of a correct answer and the ability of the candidate.

multiple-choice test on internal medicine cannot be reliable. Large samples of subjective judgements, on the other hand, can be perfectly reliable.

Recommendation 19: In ensuring reliability or sufficient universe representation, good sampling is essential. Structuring the assessment, making it more objective can help but is secondary to universe representation. In the study and in the instruments used the researcher must ensure that the sampling is sufficiently large and varied.

Cost/acceptability

Research in assessment often has the tendency to focus on validity and reliability issues almost exclusively. There are

many more issues, however, that may have nothing to do with the measurement properties of instruments. They include political and legal issues surrounding the assessment programme, technical support issues, documenting and publishing the assessment programme, R&D approaches, change management, audit methods, cost-effectiveness, accountability issues, etc. (Van der Vleuten 1996; Dijkstra et al. 2010). These elements are not trivial.

For example, good research into stakeholder acceptability is necessary because current assessment instruments rely heavily on human observation and judgement. Whereas in structured and standardised tests (for example, a multiple-choice test) reliability and validity can be built into the test paper (and it really does not matter who delivers the paper) such qualities have to be built into the user in observation-based tests. In fact, in the latter case, the 'paper part' of the assessment only serves to support and document; the actual assessment is the process between observer and student. Quality of the assessment procedure then comes from teacher training, feedback on performance, etc. If the stakeholders are not convinced about the added value of the assessment procedure and are not well instructed to use it, the results can never be valid or reliable.

Research in assessment, therefore, should pay more attention to the user and how s/he uses the instrument and the way in which the user was professionalised with respect to the assessment procedure.

Recommendation 20: Researchers should also consider topics that pertain to the embedding of the assessment within the organisation, assessment as a programme and concerning the users of the assessment to fill paucity in the literature

Ethical issues

In the case of all types of human research, it is true that certain minimum ethical standards have to be adhered to, but for research involving assessment this is even more so. Assessment is for both students and faculty an issue of high importance. We acknowledge that different countries have different procedures regarding ethical consent. In some countries, ethical committees rule educational research automatically as exempt, whereas in other countries, often a full ethical review is needed (sometimes even by medical ethical committees). Still, especially in countries where ethical review of education research is not institutionalised, the onus is on the researcher to ensure that adherence to minimum ethical standards is maintained, even if these may not have a legal status in some countries.

Recommendation 21: When an ethical review committee exists with sufficient knowledge and jurisdiction to judge the ethical stratus of a research project it should be consulted.

Recommendation 22: If there is no suitable ethical committee to submit the research proposal to, the researcher should provide information as to the ethical care taken in the research project. S/he should describe and ensure minimally the informed consent procedure, ensure completely voluntary participation, provide a correct briefing of the participants, ensure maximum avoidance of disinformation unless there is a good debriefing, utmost prevention of any physical or psychological harm to the participant, and ensured anonymity for all participants in the reports/publication.

Infrastructure and support

Much of what was said above is about things the individual researchers can and should control. However, we also want to take some positions and make recommendations about the context in which the researcher works, its enabling and boundary conditions.

The research community

The medical assessment research community can best be characterised as an open, collegial and collaboration-orientated group. We think that this is one of the success factors, one of the reasons why medical education as a scientific discipline is evolving so rapidly. International conferences that used to be visited by only 300–400 delegates nowadays easily reach numbers of over 1500. Such an environment provides a unique opportunity for cross-institutional collaboration. Such collaboration can help to improve the quality of research, because it always forces the researchers to think beyond their local problems and formulate their research questions more generically, it gives input to ideas from various angles and it can produce research with in-built replication to other contexts. Of course, it may also help boost the numbers and therefore representativeness of the results.

Recommendation 23: Whenever possible, cross-institutional research should be attempted, or at least sharing of materials and expertise should be done from an open, collegial standpoint.

The scientific journals

Scientific journals play a role in promoting the quality of assessment research, by providing the opportunity to researchers to publish their work. In the past, some restrictions have been necessary to cope with the large numbers of submission and to avoid unacceptable publication lags. Some journals have, for example, used word count limits. Fortunately, now with online publications of papers or online publications of auxiliary material (appendices, tables, etc.), a word count limit is not longer necessary for logistical reasons. Practically, all journals have abolished such limits by now. Still, we want to formulate a recommendation on this issue.

Recommendation 24: Journals should not instil word count limits for logistical reasons, but should evaluate whether the length of the paper is appropriate for the message it contains.

Terminology and its use is still a problem in assessment research. As stated in the introduction to this article, the boundaries between education and assessment and between assessment and evaluation are fading. For example, the term 'audit' probably has over 10 different meanings. Especially with key words and titles, the somewhat liberal use of terms may make it increasingly difficult for research to conduct a thorough literature search to find suitable sources.

Recommendation 25: The discipline of research in assessment (or more broadly health sciences education research) has need of a fixed taxonomy of term, equivalent to MESH headings. The scientific journals are invited to take the lead in this.

Assessment research or health sciences education research in general is not a scientific domain that easily finds its way to funding agencies. In some situations, they do not belong to the target areas of funding agencies for biomedical research because of the focus on education, nor do they belong to the domains of educational funding agencies because the research is to domain specific. We, as a research community, should join efforts in making funding agencies more aware of the relevance, the importance and the rigour of assessment research.

Recommendation 26: The assessment research community (and the health sciences educational committee) should join forces in making funding agencies more open to funding of educational research. We suggest that this should be led by the major medical education associations.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

Notes

1. Ideographic refers to a description of a specific incidental, contingent and sometimes subjective phenomenon.

2. Nomothetic refers to the generalisable aspects of the observed or studied phenomenon. The epitome of this being universal laws, such as exist in physics. Where ideographic is sometimes seen as referring to the subjective aspects, nomothetic is seen as the objective elements.

Notes on Contributors

LAMBERT SCHUWIRTH, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

JERRY A. COLLIVER, Department of Medical Education, Southern Illinois University School of Medicine USA.

CEES VAN DER VLEUTEN, Department of Educational Development and Research, Faculty of Health, Medicine and Life Sciences, Maastricht University, the Netherlands.

DAVID B SWANSON, National Board of Medical Examiners, Philadelphia, USA.

LARRY D. GRUPPEN, Department of Medical Education, University of Michigan Medical School, USA.

CLARENCE D. KREITER, Department of Family Medicine, University of Iowa Carver College of Medicine, USA.

HIROTAKA ONISHI, International Research Center for Medical Education, University of Tokyo, Japan.

LOUIS N. PANGARO, Department of Medicine, (MED)Uniformed Services University, USA.

MICHAELA WAGNER-MENGHIN, Department for Medical Education, Medical University of Vienna, Austria.

CHARLOTTE RINGSTED, Denmark Rigshospitalet, Denmark.

STEWART MENNIN, University of New Mexico School of Medicine, Brasil.

References

Bligh J. 2003. Nothing is but what is not. Med Educ 37:184–185.

Cronbach L, Shavelson RJ. 2004. My current thoughts on coeffecient alpha and successor procedures. Educ Psychol Meas 64(3):391–418.

Dawes RM, Faust D, Meehl PE. 1989. Clinical versus actuarial judgment. Science 243(4899):1668–1674.

Dijkstra J, Van der Vleuten CPM, Schuwirth LWT. 2010. A new framework for designing programmes of assessment. Adv Health Sci Educ 1537939.

Eva K, Rosenfeld J, Reiter H, Norman G. 2004. An admissions OSCE: The multiple mini-interview. Med Educ 38(3):314–326.

Harden RM, Gleeson FA. 1979. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 13(1):41–54.

Kane MT. 2006. Validation. In: Brennan RL, editor. Educational measurement., Vol. 1. Westport: ACE/Praeger. pp 17–64.

Klein G. 2008. Naturalistic decision making. Hum Factors 50(3):456–460.

McCall W. 1920. A new kind of school examination. J Educ Res 1:33–46.

Messick S. 1994. The interplay of evidence and consequences in the validation of performance assessments. Educ Res 23(2):13–23.

Miscellaneous authors 2001. Review criteria. Acad Med 76:922–951.

Miser WF. 2005. Educational research – to IRB or not to IRB? Fam Med 37(3):168–183.

Norcini J, Blank LL, Arnold GK, Kimball HR. 1995. The Mini-CEX (Clinical Evaluation Exercise): A preliminary investigation. Ann Inter Med 123(10):795–799.

Norman G. 2003. RCT = results confounded and trivial: The perils of grand educational experiments. Med Educ 37:582–584.

Norman G, Swanson D, Case S. 1996. Conceptual and methodology issues in studies comparing assessment formats, issues in comparing item formats. Teach Learn Med 8(4):208–216.

Norman G, Tugwell P, Feightner J, Muzzin L, Jacoby L. 1985. Knowledge and clinical problem-solving. Med Educ 19:344–356.

Plous S. 1993. The psychology of judgment and decision makiing. New Jersey: McGraw-Hill Inc.

Regehr G. 2010. It's NOT rocket science: Rethinking our metaphors for research in health professions education. Med Educ 44(1):31–39.

Schuwirth LWT, Van der Vleuten CPM, Donkers HHLM. 1996. A closer look at cueing effects in multiple-choice questions. Med Educ 30:44–49.

Schuwirth LWT, Verheggen MM, Van der Vleuten CPM, Boshuizen HPA, Dinant GJ. 2001. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ 35(4):348–356.

Shepard L. 2009. The role of assessment in a learning culture. Educ Res 29(7):4–14.

Swanson DB. 1987. A measurement framework for performance-based tests. In: Hart I, Harden R, editors. Further developments in assessing clinical competence. Montreal: Can-Heal Publications. pp 13–45.

Torgerson CJ. 2002. Educational research and randomised trials. Med Educ 36:1002–1003.

Van der Vleuten CPM. 1996. The assessment of professional competence: Developments, research and practical implications. Adv Health Sci Educ 1(1):41–67.

Van Merrienboer J, Sweller J. 2005. Cognitive load theory and complex learning: Recent developments and future directions. Educ Psychol Rev 17(2):147–177.

Verhoeven BH, Verwijnen GM, Muijtjens AMM, Scherpbier AJJA, Van der Vleuten CPM. 2002. Panel expertise for an Angoff standard setting procedure in progress testing: Item writers compared to recently graduated students. Med Educ 36:860–867.