General considerations regarding assessment of professional behaviour

Walther N.K.A. van Mook a,b,⁎, Scheltus J. van Luijk c, Helen O’Sullivan d, Valerie Wass e, Lambert W. Schuwirth f, Cees P.M. van der Vleuten f

a Department of Intensive Care and Internal Medicine, Maastricht University Medical Centre, Maastricht, The Netherlands
b Department of Medical Education Development and Research, Maastricht University, Faculty of Health, Medicine, and Life Sciences, Maastricht, The Netherlands
c Free University of Amsterdam, Institute for Education and Training, VU Medical Centre, Amsterdam, The Netherlands
d School of Medical Education, University of Liverpool, Liverpool, United Kingdom
e Community based Medical Education, Manchester Medical School, University of Manchester, Manchester, United Kingdom
f Department of Medical Education Development and Research, Maastricht University, Faculty of Health, Medicine and Life Sciences, Maastricht, The Netherlands

ARTICLE INFO

Article history:
Received 31 October 2008
Accepted 16 November 2008
Available online 24 December 2008

Keywords:
Problem based learning
Professional behaviour
Professionalism
Tutorial group
Assessment

ABSTRACT

Core medical knowledge has been assessed for over fifty years and technical and communication skills for at least twenty. The assessment of professionalism however has failed to achieve recognition within this time frame. The interest in the assessment of professionalism and professional behaviour thus is a fairly recent development. This article will firstly clarify how professional behaviour assessment relates to other assessment methods using the framework proposed by Miller. Thereafter a brief overview will be provided of the current “tool box” of methods available to assess professionalism. Data on the validity, reliability, feasibility, acceptability and educational utility of these “tools” as derived from published evidence will be reviewed. Subsequently a general overview of the way forward in the assessment of professionalism and professional behaviour will be given.

© 2008 European Federation of Internal Medicine. Published by Elsevier B.V. All rights reserved.

1. Introduction

Core medical knowledge has been assessed for over fifty years and technical and communication skills for at least twenty. The assessment of professionalism however has failed to achieve recognition within this time frame [1]. This has sent a conflicting message to both students and practicing physicians [2]. Early interest in the assessment of professionalism focused on the process of professional development and the attainment of “professional identity” [3]. Later studies concentrated on identifying things that were wrong, or missing, such as loss of ethical principles or idealism, a so called “deficit” model [3]. Truly professional physicians, however, not only avoid doing harm (nonmaleficence). They also seek to do good (beneficence). Today we see professionalism not as the absence of negative attributes, but as a set of identifiable positive qualities or behaviours that everyone should aim for [3]. In a competence based curriculum, it is particularly important to set the expectation that students develop excellence in professionalism. It has become evident that to do this providing experiences and setting expectations for teaching and learning professionalism on their own are not enough. Professionalism must be assessed if it is to be viewed as both positive and relevant [4].

This interest in the direct assessment of professionalism is thus a fairly recent development. The challenge of assessment of professional behaviour can be summarized in two statements: “If it can’t be measured, it can’t be improved”, and “They don’t respect what you expect, whereas they respect what you inspect” [5]. Assessment thus is a mechanism by which medical faculties signal that they value certain subject areas. The lack of formal assessment may undermine the impact of teaching. Therefore assessments must be established to measure the requisite attributes of professionalism. Unfortunately professionalism and, to a somewhat lesser extent, professional behaviours, are difficult competency domains to assess.

This article will firstly clarify how professional behaviour assessment relates to other assessment methods using the framework proposed by Miller [6]. Thereafter a brief overview will be provided of the current “tool box” of methods available to assess professionalism. Data on the validity, reliability, feasibility, acceptability and educational utility of these “tools” as derived from published evidence will be reviewed. Subsequently a general overview of the way forward in the assessment of professionalism and professional behaviour will be given.

2. Professional behaviour assessment in Miller’s pyramid

The framework for the assessment of clinical skills, competence and performance as proposed by Miller [6] is often used to illustrate
the relative position and use of commonly assessment instruments in medical education (see Fig. 1). Although many variants exist, Miller’s four levels of hierarchy form the basis for interpreting the pyramid’s structure. The lowest two levels (knows, knows how) test aspects of cognitive knowledge. The upper two levels (shows and does) focus on behavioural aspects. The lower levels therefore play a relatively important role in the early phases of medical education. The upper levels are more important in the clinical undergraduate phases and postgraduate training. The context authenticity increases with each level. Workplace based assessment forms the summit of the pyramid and aims to measure performance in everyday practice. ‘Does’ can be viewed as synonymous with ‘performance in vivo’, and ‘shows how’ as a measure of ‘performance in vitro’, or “competence”. It should however be realised that the ‘workplace’ of first and second year undergraduate students is often formed by small (tutorial) group sessions, in contrast to the clinical workplace of later years and in postgraduate training. The upper two levels are most frequently assessed by direct observation. This is not a prerequisite for the lower levels. Professional behaviour assessment aims at measuring the ‘does’ level of Miller’s pyramid, i.e. actual performance in daily practice.

3. A brief overview of the professionalism tool box

At least 88 different methods for assessing professionalism have become established in medical education between 1982 and 2002 [7]. Since then several have been added [2]. Currently the most commonly used approaches are peer assessment, the objective structured clinical examination (OSCE), direct observation of behaviour by faculty members, critical incident reports and learner maintained portfolios [5,8,9]. Standardised checklists, written comments and reports from formal evaluation sessions completed by a supervisor and/or other staff are also commonly used [10,11]. Other methods of assessing professional behaviour include short written reflective narratives (about important incidents) [12], and self-assessment [13].

These tools for assessing professional behaviour should meet the criteria of validity, reliability, feasibility (or which cost is an important element) and acceptability (to all stakeholders) [14–17]. The criteria are shown in Table 1. Apart from these elements, the impact on education and learning is of importance [14]. Preferably, all assessment methods should have these criteria. However the ideal world is a utopia. In daily practice a trade-off exists between these utility criteria [17]. The ‘utility’ or usefulness of an instrument is a function of the relationship between several elements. The cost aspect of feasibility has an inverse relationship with practical utility [14,17]. In practice compromise is inevitable. Different weighting must be applied to these criteria in different individual situations depending on the context and purpose of the assessment. In a high stake assessment, reliability will have higher priority in the choice of assessment method. In the context of in-training assessment, where the final decision is based on a triangulation of different assessments, reliability can be compromised in favour of educational impact [14].

The application of the utility equation to the assessment of professionalism is discussed in the following paragraph.

4. Data available from the literature

Several reviews on the assessment of professionalism have been published. Different approaches have been used and different conclusions reached [7,16–20]. Four of these reviews have been selected for discussion.

Arnold classified assessment methods, derived from 170 papers over a 30-year period, into three groups. These addressed professionalism as either (i) part of clinical competence or, (ii) a single construct or (iii) as separate elements of professionalism (e.g. humanism or ethical reasoning) [17,18]. She stressed the need for the development of rigorous qualitative methods to strengthen pre-existing quantitative methods, embrace high order behaviours (such as the resolution of conflict), and to acknowledge the context in which the professional behaviour occurs [18].

Ginsburg’s review delved deeper into the conceptual issues of professionalism [18,19]. The authors concluded that future work on understanding and evaluating professionalism, should focus on behaviours not personality traits or vague concepts of character. The assessments should furthermore include context and conflict to ensure they were relevant and valid. In their view, peer assessment and self-assessment are also essential components as they are increasingly being used in the continuing professional development of all practicing physicians. Finally, they recommend that assessors should attempt to understand why students occasional have lapses in professional behaviour, in order to develop effective teaching and remediation in this domain [19].

Lynch reviewed 191 articles published over a 20 year period and identified 88 instruments available (at that time) for the assessment of professionalism for education or research purposes [7]. Most looked at ethical reasoning, and only 27 were described as comprehensive (i.e. measuring two or more elements of professionalism). Most methods were based on self-assessment and delayed recall [7,17]. Information on the reliability and validity of the methods varied considerably. The authors concluded that the focus should now be on improving existing assessments rather than developing new ones. In addition more research was needed on predictive validity and tailoring the assessments for longitudinal use. They argue we need to understand more about how the environment in which these assessments take place influences the outcome. Assessment should be more performance-based and longitudinal [7,17].

Epstein and Hundert confirmed that assessment methods rarely relied on observations in daily practice and real-life situations or incorporated views of either peers or patients. Nor did they use measures that predicted clinical outcome [17,20]. They highlight the need to develop new tools, as well as refining existing ones, the ability to handle uncertainty and work in a team [20]. Perhaps their most important recommendation is that an inadequate system of feedback, mentoring and remediation will subvert even the most well-conceived and validated examination [20].

Finally, the most recent review by Veloski reported on 134 studies on professionalism. Approximately half attempted to estimate reliability but 72 gave no information at all. Content validity (the extent to which a measure represents all facets of a concept) was mentioned in 86 studies but only addressed in detail in 34. Construct validity (the extent to which a measure represents a certain psychological construct) was discussed in 61 studies. Some evidence of practicality was reported in about two thirds of studies [16] but again detailed information on practicality can only be found in a third.

Fig. 1. The framework for assessing competence as proposed by Miller: Miller’s pyramid [6] Abbreviations: MCQ = multiple choice questions OSCE = objective structured clinical examination.
Assessment may drive learning through content: tasks should reflect professional or educational reality as closely as possible.

- Reliability
  - The setting must allow a judge to be able to observe the student several times repeatedly over a longer period of time and preferably in many different contexts (intra-observer reliability, inter-cases reliability, and context specificity).
  - The prevalence of unprofessional behaviour in an institution influences the reliability of the assessment.
  - Assessors must be trained to use and interpret the rating scale.
  - The overall assessment of the student is constituted from independent observations of different assessors.

- Validity
  - The setting must allow a judge to be able to observe the student several times repeatedly over a longer period of time and preferably in many different contexts (intra-observer reliability, inter-cases reliability, and context specificity).
  - The prevalence of unprofessional behaviour in an institution influences the reliability of the assessment.
  - Assessors must be trained to use and interpret the rating scale.
  - The overall assessment of the student is constituted from independent observations of different assessors.

- Acceptability
  - The student must have and be given the time to improve his/her professional behaviour.
  - The method and its educational background should be clear to the assessor.
  - Information should be provided to get commitment and willingness to compromise.

- Feasibility
  - Assessors must be trained to use and interpret the rating scale.
  - The setting must allow a judge to be able to observe the student several times repeatedly over a longer period of time and preferably in many different contexts (intra-observer reliability, inter-cases reliability, and context specificity).
  - The prevalence of unprofessional behaviour in an institution influences the reliability of the assessment.
  - Assessors must be trained to use and interpret the rating scale.
  - The overall assessment of the student is constituted from independent observations of different assessors.

- Educational impact
  - Assessors should constantly be aware of the driving force of assessment to achieve desirable educational effects.
  - Assessors must be trained to use and interpret the rating scale.
  - The overall assessment of the student is constituted from independent observations of different assessors.

Only eleven studies addressed professionalism as an integral construct of professional behaviour, and nine as a distinct facet of clinical competence. In fact 65 of the 134 studies used self-administered instruments rather than independent observation of behaviour [16]. In most (109 studies), the tools were part of research or program evaluations. Reports on their use as formative or summative assessment were relatively uncommon.

5. Considerations regarding assessment of professionalism

As so many different methods are available for assessing professionalism, it seems self-evident that one single instrument in isolation fails to give a true measure. Combining several methods focusing on different aspects of professionalism is preferable. In other words: there is need for ‘triangulation’ [17]. Although data from the literature on the reliability and validity is currently limited, general guidelines for those assessing professional behaviour assessment emerge. They are summarized in Table 2 and will be discussed more extensively in the paragraphs below.

First, the purpose of the assessment must be clear. Does it serve a formative or normative purpose, or perhaps both? [21] Formative assessment aims to develop and steer student behaviour, whereas summative assessment may have important consequences for student progression in the curriculum. The former provides constructive specific feedback; the latter makes a pass/fail, “go-no go” decision. In the case of professional behaviour, specific and detailed feedback derived from assessment may improve professional behaviours [22,23]. By conducting assessments frequently, with long-term longitudinal follow-up, it provides the opportunity for learners to change, thereby guiding remediation [15,24]. Checklists alone have low detection indices on which to build feedback. This can improve when supervisors have the opportunity to make descriptive comments [25]. Therefore the preferable aim when assessing professional behaviour should be formative. In actuality
within medical education it is more frequently summative in nature [7].

Furthermore, the choice of outcome (e.g. selection, feedback, improvement of the educational program) is intimately linked to the utility of the assessment method. For example, peer assessment works well when used to identify good professional behaviour. When used as a summative tool to identify those in need of remediation students either don’t participate or they conspire against the system to provide only excellent assessments of peers [26].

Two additional elements influence the observations of assessed behaviour: transparency (disclosure of the process and outcome of the assessment), and symmetry (all levels in the organizational hierarchy are evaluated using the same methods). To ensure transparency the purpose of the professionalism assessment must also be clear. Symmetry is an aspect frequently sought by medical students [27]. To our knowledge, 360° assessment of professional behaviour to include assessment of clinicians and other staff is not routine practice in many medical schools.

Subsequently it is important to decide what criteria to set for the professionalism assessment. Should it be evaluated from within the context of a reference group, or via preset standards for professionalism generated, for example, by an expert panel [28]. The first is termed norm referencing, the second criterion referencing. It can be envisaged that when setting norm-referenced standards any group will be initially relatively unfamiliar with the topics of professionalism under assessment. It can only be useful, once the informal standards or definitions have been shaped by experience, norm-formation, and anchorage of these norms within group culture (the assessors have become experts in the field) [28]. Until this is the case, criteria referenced standards are preferable [28]. For both methods, whether it be norms or criteria, setting the reference standard remains difficult.

What should we use as a gold standard to define the presence of optimum professional behaviour? As in clinical medicine sensitivity and specificity never attain 100% (the perfect test is fiction), clinicians tend to look at the test’s predictive value. The prevalence of a disorder in the population studied (or prevalence of unprofessional behaviour or professional lapses among medical students) is of paramount importance when attempting to answer this question. In a population of students with a low incidence of professional lapses, as for example determined for the Faculty of Medicine in Maastricht, Netherlands, the chance of a false negative test (a student with unprofessional behaviour that is not detected by a test) increases. As illustrated by the influence of prevalence of a disorder or disease in the population under study, test results (for professionalism) are thus also institution dependent [28]. Increasing the number of tests (extending it to a more wide range of relevant situations) will increase the reliability of the result.

The reliability of global rating scales can be increased by using anchor statements and by training the assessors [17]. The differing stringency with which evaluators judge, even when observing the same performance, is a well known fact and problem [29]. Apart from training the assessors, judgement of a student’s professional behaviour by several different teachers is another necessary prerequisite to protect students and staff against hasty conclusions [15]. Teachers tend to make judgments based on limited observations. They over-generalize and introduce “attribution bias” i.e. the tendency to assume that if people behave badly in one situation, they will behave badly in all situations [2]. Different assessors offer different perspectives, thereby enhancing the breadth of assessment and its reliability [30]. In other words, several faculty members are needed to produce reliable results (thereby decreasing the inter-observer variability) [31]. Increasing the number of tests in a range of relevant contexts (see below) as well as the number of different observers will thus increase the reliability of the assessment. The Dutch Project group Consilium Abeundi reports that reliable judgment depends on at least 4 to 6 judgments by different, independent assessors [32].

The context in which observations are made should be as realistic as possible. The closer the context of assessment to reality, the more valid the evaluation of professionalism is likely to be. The assessment should preferably aim at assessing performance (the way a person behaves in everyday practice) rather than competence (the behaviour displayed under test conditions). There may however sometimes be gaps between observed performance and competence. Performance may exceed competence (photographic memory but poor critical judgment), and vice versa (test anxiety but excellent clinical judgment) [20]. The assessor has therefore to establish whether the observations are representative. In attempts to measure performance in the most authentic context possible, assessment should preferably be performed in the workplace [17].

For an authentic judgment of professionalism, several authors advise that the assessment should ideally include a situation involving conflict [2,18,19]. Apart from solving the problem presented, the reasoning behind the resolution of a dilemma is important [2]. Additionally, the assessment of professional behaviour should take place over time and incorporate a longitudinal framework. Professionalism is not static, it is developmental and subject to change [28,33,34]. To provide such a longitudinal trajectory with adequate follow up and opportunities for remediation, behaviour has to be observed from the start of the program. Currently assessment of medical student professionalism is often delayed until clerkships and internships are commenced [22,35]. The challenge is to optimise the conditions for the acquisition and maintenance of professionalism in medicine, by providing stage-appropriate experiences to students [33,34]. In-course assessment, by providing a several measures over time, arguably offers a fairer view of students’ professional conduct than end-of-course assessment alone [36].

Assessment and judgment of professional behaviour aim to assess observable behaviour itself on the one hand, and the ability to reflect and feedback on that behaviour on the other. Reflection on clinical encounters, and following critical incidents, is important in both daily clinical practice as well as personal development [17]. Reflection contributes to identity and character development, and individual learning [37]. The ability to assess the reflective and feedback capabilities of a student depends on the context and assessment method used as well as the phase of training of the trainees. For example, using a scoring form in the tutorial groups to judge a student’s ability to reflect and respond to feedback on professional behaviour proved difficult in the first two years of medical school [21]. In all phases of the curriculum, intensive guidance by mentors, for example using a portfolio, can stimulate self-awareness by enabling
students to critically examine their own professional behaviour [38]. Providing constructive feedback using qualitative information gathered during such assessments has been shown to result in improvement of professional behaviour [22,23].

It is also recommended that judgments should not rely on delayed recall, since use of assessment methods that do (for example some rating scales) can result in decreased validity and reliability [39]. Several publications however report professional behaviour assessment methods that involved data collection concurrent with or immediately following observations [40–44].

Finally, several studies published over recent years have examined the extent to which assessment data on professionalism related to future professional performance [7,10,22,23,45–47]. Those who were described as “irresponsible” or as having “diminished ability to improve their unprofessional behaviour during medical training” are those who are most likely to experience disciplinary action during active practice as a physician [46]. Conscientious behaviour (as measured by immunisation and course evaluation compliance) in pre-clinical years was found to be predictive of outcomes in professionalism as found by the review board in clinical setting in year 3 [48]. Similar findings were recently reported for residents in internal medicine [47]. Therefore, once professional behaviour has been observed and adequately described, a positive or negative judgment needs to be reached and consequent outcomes attributed to this judgment. This may not as simple as it sounds. When a student is ‘academically’ competent (sufficient marks for technical skills and knowledge), but persistently fails his/her professionalism assessment it can be difficult to implement. The approach to the student displaying unprofessional behaviour will be subject of one of the forthcoming issues of the European Journal of Internal Medicine.

6. Conclusions

Assessment of professional behaviour should meet several requirements. In general there must be clarity about the purpose of the assessment, the formative or summative character and its consequences. The assessment should aim at assessing performance rather than competence, by assessing using multiple methods (triangulation), in varied contexts, preferably as realistic as possible, with multiple observers over a longitudinal trajectory with adequate follow-up, and with opportunities for remediation. Complying with these requisites increases the likelihood that the assessment will be accurate (reliable), and appropriate (valid). It should be realised however that the choice of outcome contributes to the utility of the assessment tool as well.

In the assessment of professional behaviour, the importance of reflection and feedback to promote learning should not be ignored. When feedback and opportunities for remediation of poor professional behaviour prove fruitless, it needs to be decided whether the observations warrant a negative summative judgment with punitive consequences. Although many different approaches to professional behaviour assessment have been published, content validity, reliability, and feasibility have been insufficiently studied for most instruments. However, although it needs to recognised that they are not “state-of-the-art tools”, combinations of the currently available methods, will have to suffice, until more reliable and valid instruments become available.

7. Learning points

• The ‘magic bullet’ regarding professionalism assessment does not exist.

• Current ‘state-of-the-art’ in the assessment of professional behaviour is combining multiple methods (triangulation) in varied contexts with multiple observers over a longitudinal trajectory.

• Current state of the art in education of professional behaviour is assessment with adequate follow-up and with the provision of opportunities for remediation; only assessment without remediation loses ground.

References


[10] Van Luijk SJ, Smeets SGE, Smits J, Wolfhagen IH, Perquin MLF. Assessing professional behaviour during medical training is combining multiple methods (triangulation) in varied contexts with adequate follow-up, and with the provision of opportunities for remediation; only assessment without remediation loses ground.


[20] Van Luijk SJ, Smeets SGE, Smits J, Wolfhagen IH, Perquin MLF. Assessing professional behaviour during medical training is combining multiple methods (triangulation) in varied contexts with adequate follow-up, and with the provision of opportunities for remediation; only assessment without remediation loses ground.


