Assessing professionalism: a review of the literature

DEIRDRE C. LYNCH1, PATRICIA M. SURDYK1 & ARNOLD R. EISER2
1 Accreditation Council for Graduate Medical Education, Chicago, IL, USA; 2 Drexel University College of Medicine and Consultant to the American Board of Medical Specialties, Philadelphia, USA

SUMMARY Although assessing professionalism poses many challenges, gauging and detecting changes in professionalism is impossible without measurement. This paper is a review of techniques used to assess professionalism during the past 20 years. The authors searched five electronic databases and reference lists from 1982 to 2002. Eighty-eight assessments were retained and organized into content area addressed (i.e., ethics, personal characteristics, comprehensive professionalism, diversity) and type of outcome examined (i.e., affective, cognitive, behavioral, environmental). Instead of creating new professionalism assessments, existing assessments should be improved. Also, more studies on the predictive validity of assessments and their use as part of formative evaluation systems are recommended. Based on the review, suggestions are presented for assessing medical students, resident physicians and practicing physicians.

Introduction

Society expects physicians to act professionally. In response to recent criticism regarding unprofessional behavior in medicine, some argue that improving medical professionalism can only occur through changes in teaching and assessing it (Crues, 1997; Relman, 1998). Although assessing professionalism poses many challenges, gauging and ascertaining growth in professionalism is impossible without measurement. Concluding her review of approaches to assess professionalism in medical education, Arnold stated that “[w]ithout solid assessment tools, questions about the efficacy of approaches to educating learners about professional behavior will not be effectively answered” (Arnold, 2002).

The present review defines medical professionalism as the ability to meet the relationship-centered expectations required to practice medicine competently. It varies along a continuum from positive to negative engagement and encompasses constructs such as: respect for others, integrity, altruism, accountability, duty, composure and sensitivity to diversity (ACGME, 1999; ABIM, ACP-ASIM, & EFIM, 2002; Kuczewski et al., 2003; Surdyk et al., 2003).

Two reviews, that by Arnold (2002) cited above and the other by Ginsburg & colleagues (2000) provide the most recent analyses regarding conceptual issues pertinent to assessing professionalism. Arnold categorized instruments into three groups: those assessing professionalism as part of clinical performance; those assessing it as a comprehensive entity; and those assessing separate elements of professionalism, such as humanism and ethical decision-making. She suggested that assessments could be improved by focusing on professionalism as a separate entity, including rigorous qualitative approaches, exploring the extent to which the environment supports assessment, and determining the usefulness of developmental expectations in assessing professionalism. In their review, Ginsburg and colleagues (2000) examined existing assessments in terms of types of raters, such as faculty supervisors and peers. They reasoned that difficulties involved in assessing professionalism are due to the frequent use of abstract idealized definitions, the context-specific nature of professionalism, and evaluator reluctance to address relatively minor lapses. The same authors proposed that professionalism assessments should address cognitive and behavioral outcomes.

The purpose of the present review is twofold: (1) to extend and update the work of previous authors, and (2) different from the format of previous reviews, to provide a catalog of the many and varied assessments available in the literature. The first goal is achieved by including assessments that measure professionalism in the educational environment and assessments that address diversity. The second goal is achieved by: (a) organizing assessments according to content and to types of outcomes assessed, and (b) listing assessments in a table together with brief descriptive summaries for easy reference by educators searching for specific assessments. The concluding section presents general recommendations for designing and implementing a system to assess professionalism and specific suggestions for assessing medical students, residents and practicing physicians.

Method

Data sources

Two authors independently searched five electronic databases, from 1982 to 2002: MEDLINE, the Educational Resources Information Center (ERIC), Topics in Medical Education Literature (TIMEMLT), Health and Psychosocial Instruments (HAPI), and PsychINFO. Twenty-eight search terms included: professionalism, duty, ethics and variations of these terms or their combination with assessment, measurement, and update the work of previous authors, and (2) different from

Correspondence: Deirdre C. Lynch, Accreditation Council for Graduate Medical Education, 515 N. State St., Suite 2000, Chicago, IL 60610-4322, USA. Tel: (312) 755 7441; fax: (312) 755 7498; email: dlynch@acgme.org

DOI: 10.1080/01421590410001696434

366 ISSN 0142-159X print/ISSN 1466-187X online/03/030366-8 © 2004 Taylor & Francis Ltd

designed to address professionalism in conjunction with other competence areas were included if the professionalism items comprised a defined subscale or scale of more than one item or a defined set of qualitative categories. Excluded were assessments that measured learner or teacher course satisfaction as well as assessments that addressed specific medical conditions (e.g. attitudes toward people with HIV) or specific ethical issues (e.g. advance directives).

Procedure

The reviewers pooled their findings to create a single list of published articles. This list was divided into three sets; each set was assigned to a reviewer who identified approaches used to assess professionalism and extracted key information about each. Two experts in professionalism reviewed the final list to confirm that recognized professionalism assessments had been included.

To ease reference, the assessments were categorized according to four content areas addressed (ethics, personal characteristics, comprehensive professionalism and diversity) and four types of outcomes examined (affective, cognitive, behavioral and environmental). This resulted in 16 possible categories for organizing the assessments. Information gathered about each assessment followed the format used by Boon & Stewart (1998) in a similar review of communication assessments. This format includes items that constitute the headings in Table 1 described below. (Note: Tables 1a and 1b are too long to be included here and are available on the Medical Teacher website: www.medicalteacher.org)

Name of the assessment

The first column of Table 1 presents the name of the assessment instrument or approach. Assessments without specific names were assigned a brief descriptive title using the first author’s name. The reference for each assessment is in parentheses.

Description

Column two describes assessment approaches in terms of what they assessed (e.g. opinions about professional behaviors), how data were collected (e.g. self-administered rating form) and the kind of data collected (e.g. quantitative using seven-point scale, qualitative). Column three describes the number of assessment items or the number of categories used to code qualitative data.

Reliability type and coefficient (a number between 0 and 1) are recorded in column four. Reliability refers to the consistency of assessment results (AERA, APA, NCME, 1999). Inter-rater or inter-observer reliability is the extent of agreement across two or more observers rating the same performance. Inter-item (e.g. Cronbach’s alpha), inter-station, or inter-case reliability is the degree of consistency in an individual’s performance across different items, stations or cases. Test–retest reliability is an indicator of consistency over time. Generalizability theory offers an alternative approach to assessing the individual reliabilities listed above. Variables that might affect reliability are examined simultaneously, instead of separately, thus providing an overall reliability index termed a G coefficient. With regard to qualitative data, inter-rater agreement on coding is important. When information about reliability was not included in the article, ‘none reported’ is noted.

Evidence used to infer validity

Column five reports information about validity. Validity is the extent to which an assessment approach measures what it intends to measure (AERA, APA, NCME, 1999). There are different types of validity and the evidence used to infer each may be obtained in several ways. Content validity is often determined by expert review of the assessment items to see if they adequately represent the domain being assessed. Construct validity may be examined by conducting a factor analysis, by seeing if the assessment discrimiates among levels of trainee (e.g. third-year residents perform significantly better than first-year residents), or by determining whether the assessment is responsive to learner changes (e.g. those who complete an ethics course perform better at the end of the course than at the beginning of the course). Criterion validity is inferred when assessment results correlate with data from other assessments designed to measure the same thing (concurrent validity) or with future indicators (predictive validity). Finally, face validity refers to the extent to which an assessment measures what it appears to measure and may be ascertained via pilot testing or expert review.

With qualitative data, validity may be inferred by triangulation whereby information from two or more sources is examined to determine similarities among the sources (Strauss & Corbin, 1998). In addition, asking participants to check the accuracy of transcripts generated from focus groups or interviews may validate qualitative data. When information about validity was not included in the article, ‘none reported’ is noted.

Notes

The last column describes the purpose of the assessment, as either research or as part of an educational process; setting and participants are also described. It also notes characteristics of the assessment likely to influence feasibility. Examples of these characteristics include special training, technical equipment, extensive time requirements and cost.

Results

Of approximately 6200 abstracts, 359 were selected for further review and manual reference list searches yielded 52 more references. Of these 411 articles, 220 were excluded because they did not meet selection criteria. The remaining 191 articles provided descriptions of 88 assessments (Table 1a) [1] that fit into one or more of the following 12 categories: ethics–affective (15), ethics–cognitive (25), ethics–behavioral (9), ethics–environmental (14), personal

Ethics
Forty-nine assessments addressed various aspects of ethics such as morality, ethical principles, honor codes, social norms, deception, abuse or mistreatment, cheating, disclosure, and sexual misconduct. Validity and reliability were examined in only nine of these assessments (Table 1b). Of these, cognitive outcomes were measured by the following: the Christie Ethical Decision-making Questionnaire (Hoffmaster et al., 1991), the Defining Issues Test (Self et al., 1992), the Professional Decisions Values Test (Rezler et al., 1992) the Savulescu Ethics Competence Tool (Savulescu et al., 1999), the Siegler Assessment (Siegler et al., 1982), the Sulmasy Questionnaire for House Officers (Sulmasy et al., 1995), and the Wenger Orthopaedic Surgeon’s Knowledge of Medical Ethics Questionnaire (Wenger & Lieberman, 1998). Behavioral outcomes were measured in two assessments: the Ethics OSCE (Singer et al., 1996) and the Moral Behavior Analysis (Shechan et al., 1987). Either validity or reliability was examined in nine assessments designed to measure affective outcomes and the educational environment. Of these, the Levitt Ethical Issues Questionnaire (Levitt et al., 1994), which provided evidence for face and content validities and has been used with residents and faculty physicians, appears to be the most efficient approach to measuring affective and environmental outcomes.

Personal characteristics
The category labeled ‘personal characteristics’ included assessments that addressed attributes such as emotional intelligence, personal values, empathy, trustworthiness, cynicism and dogmatism. All of the personal characteristics instruments assessed affective outcomes. Validity and reliability were examined for 13 of 15 (Table 1a). Of this pool, only two instruments were completed by patients, the Trust in Physician Scale (Anderson & Dedrick, 1990) and the Wake Forest Physician Trust Scale (Hall et al., 2002); the rest were self-administered. Patient input questionnaires typically yield data that are positively skewed and demonstrate ceiling effects (Matthews & Feinstein, 1989). The Wake Forest Physician Trust Scale, however, addresses these issues. When it is not feasible to obtain patient input, the Schwartz Values Scale (Eliason & Schubot, 1995) and the Jefferson Scale of Physician Empathy (Hojat et al., 2002) may be feasible alternatives to aid teaching and learning about personal meaning schemas (i.e. beliefs, values, attitudes) and their influence on professional behavior.

Comprehensive professionalism
Comprehensive professionalism pertained to assessments that addressed two or more components of professionalism in a single assessment. Validity and reliability were examined in 11 of 27 comprehensive professionalism assessments (Table 1a). Of these, the Nurse Evaluation of Medical Housestaff Form (Butterfield & Mazzaferr, 1991), the Amsterdam Attitudes and Communication Scale (DeHaes et al., 2001), and the Humanism Scale (Hauck et al., 1990) appear to be the most credible instruments to measure behavioral outcomes. The Humanism Scale has been used to assess practicing physicians (Hauck et al., 1990) and medical students (Coutts & Rogers, 2000). Data yielded from it have shown significant positive relationships with adherence to medical advice and patient satisfaction (Hauck et al., 1990). Because it requires direct observation and simultaneous documentation, the Amsterdam Attitudes and Communication Scale helps to mitigate rater biases due to general impressions or recall. The Nurse Evaluation of Medical Housestaff Form is one of the few validated approaches to assess medical professionalism by non-physician co-workers. The participant observation method employed by Stern (1996) is an ideal approach to assess professionalism in the educational environment. On the other hand, the 12-item, self-administered Scale to Measure Professional Attitudes and Behaviors in Medical Education is probably more feasible (Arnold et al., 1998).

Diversity
Diversity pertained to assessments that addressed cultural issues, socioeconomic status, gender, age or disability. Validity and reliability were examined for five of 10 instruments (Table 1a). Of these, the Cultural Competence in Medicine Questionnaire (Godkin & Savageau, 2001) and the Sociocultural Attitudes in Medicine Inventory (Tang et al., 2002) were able to detect post-intervention improvements in affective outcomes among medical students. The Cultural Competence in Medicine Questionnaire also measured cognitive outcomes. One diversity assessment, the Robins Health Beliefs Communication OSCE (Robins et al., 2001), measured behavior and one assessment measured the environment, the Elam Diversity in Medical School Questionnaire (Elam et al., 2001).

Discussion
This review identified and analyzed professionalism assessments used in medical education from 1982 to 2002. Most assessments addressed medical ethics, particularly ethical knowledge and reasoning. This finding may reflect a historical emphasis on assessing ethics versus professionalism in medical education; today ethics is considered one element of professionalism. The next largest group of assessments addressed behaviors in the category of comprehensive professionalism, usually in clinical settings. It is surprising that so few assessments addressed diversity and cognitive outcomes of comprehensive professionalism, given the importance accorded these areas. The assessments demonstrated great variability in the extent to which reliability and validity were examined. Psychometrics was most likely to be examined in assessments that measured ethics knowledge or reasoning and personal characteristics. Psychometrics was least likely to be examined in assessments that addressed ethical attitudes
or beliefs, ethical behavior and ethics in the educational environment. Many assessments relied on self-reported data obtained in cross-sectional research studies. Only five comprehensive professionalism assessments that measured behaviors involved data collection concurrent with or immediately following observations (DeHaes et al., 2001; Beckman et al., 1990; Carney & Mitchell, 1986; Couts & Rogers, 2000; Prislin et al., 2001). Most relied on delayed recall of observations, which decreases data validity and reliability (Fletcher et al., 1982; Gray, 1996). Only three assessments examined the extent to which assessment data related to future professional performance (Herman et al., 1983; Phelan et al., 1993; Papadakis et al., 2001). This key shortcoming in the literature warrants further investigation. Furthermore, very few assessments were designed for, or implemented as part of, longitudinal studies of professionalism. This is another area requiring additional work, given the need for continuous and frequent assessment of professionalism and recent attention to designing assessments tailored to developmental or training levels.

**General recommendations for assessing professionalism**

Summative assessment, often undertaken to classify learners, dominates medical education. Using assessment to improve learning is gaining credibility, however, with reports of formative assessments appearing more often in the literature. Some evidence suggests that focused teaching can improve moral reasoning (Self et al., 1992; Goldie et al., 2002); more specifically, that feedback derived from assessment may improve professional behaviors (Phelan et al., 1993; Papadakis et al., 2001). These findings suggest that professionalism should be formatively assessed. This means that assessment should begin early (Lowe et al., 2001) be conducted frequently, be implemented long term, and provide learners with opportunities to change (Van Luijk et al., 2000).

The educational environment, whether through formal or informal curricula, appears to influence learner attitudes and behavior (Stern, 1996). One study discerned relationships between the ethical environment and medical students’ ethical behavior (Feudner et al., 1994). In another study, residents reported learning most about professionalism from observing role models (Brownell & Cote, 2001). Research suggests that the business (Freeman et al., 1999) and cultural environment (Hoffmaster et al., 1991) influence professionalism among practicing physicians. Consequently, assessments that gauge professionalism in the environment may provide insight into the professionalism of individuals.

Systematic assessment of professionalism should also include many different assessors, more than one assessment method and assessment in different settings (Littlefield et al., 1996; Van Luijk et al., 2000; Ginsburg et al., 2002). Research indicates that different assessors offer different perspectives, thus enhancing the breadth of assessment (Wooliscroft et al., 1994) and multiple assessors enhance reliability (Swanson, 1987). Each assessment method has strengths and weaknesses. Rating forms are considered relatively easy to use, but are plagued with the ‘halo’ or ‘horns’ effect (Gray, 1996). The Ethics OSCE may mitigate the latter weakness but requires several cases, and hence testing time, to obtain stable estimates of learner performance (Singer et al., 1996). Consequently, using more than one assessment method may help to compensate for the weaknesses associated with any single approach. Because professionalism is a complex construct, it is unlikely that a single assessment will adequately measure it. Using a combination of assessments, however, such as a moral reasoning assessment together with a behavioral assessment, may be adequate. Assessment in different settings can help to determine the generalizability of learners’ professionalism and identify context specific issues relevant to learners. Suggestions for assessing medical students, residents, and practicing physicians Medical students. Assessment of medical student professionalism is often delayed until clerkship rotations. The Physicianship Evaluation Form (Papadakis et al., 2001), however, indicates that it is both desirable and possible to begin assessing student professionalism during the first year of medical school. Initial use of this assessment is formative; students are invited to present their perspective and data are used to provide feedback and guide remediation. Persistent patterns of unprofessional behavior, despite remediation, may provide grounds for dismissal. This performance-based, longitudinal approach helps to set professionalism expectations early, both for learners and for faculty who must commit to addressing and attempting to improve students’ professional behavior. Students should also be prepared to address ethical issues in clinical settings. Behavioral simulations, such as the Moral Behavior Analysis (Sheehan et al., 1987) and the Ethics OSCE (Singer et al., 1996), in which ethical dilemmas are role-played, are ideal for this purpose, especially if they include post-encounter oral or written assessments that probe reasoning used during encounters. To obtain reliable data, many cases would be required, thus, these simulations should be used for formative assessment only.

Residents. Residents typically rotate through many different settings and interact with a broad range of personnel. These characteristics of residency training may present challenges to coordinating comprehensive assessment and providing opportunities for the development of long-term relationships helpful to assessing professionalism. On the other hand, changing rotations provide the option of obtaining input from various observers (e.g. nurses, patients, supervising physicians and peers), and assessing the extent to which resident professionalism varies across settings. Three-sixty assessments, which obtain data from key people within a learner’s sphere of influence, capitalize on the availability of a broad scope of potential assessors (Tornow & London, 1998). The Nurse Evaluation of Medical Housestaff Form (Butterfield & Mazzaferrri, 1991), which involves assessment by nursing staff, may be adapted for that purpose. To assess residents’ knowledge of professionalism and to stimulate instructional discussion, the Barry’s Challenges to Professionalism Questionnaire may be used (Barry et al., 2000). This self-administered questionnaire consists of vignettes that address conflict of interest, gifts and physician impairment. Research results indicated that the questionnaire was able to detect a broad range of knowledge about professionalism and discriminate among respondents depending on years of experience.
Assessments of ethics knowledge by surveys and of comprehensive professionalism by patient questionnaires were the most frequently used, and are probably the most feasible, approaches to assessing professionalism in practicing physicians. For instance, the Defining Issues Test, a written test in which ethical dilemmas are presented in vignettes, is capable of discriminating different levels of moral reasoning among practicing physicians (Baldwin & Bunch, 2000). The Barry’s Challenges to Professionalism Questionnaire would also be suitable for use with practicing physicians.

Examples of patient questionnaires designed to assess practicing physician professionalism include the Humanism Scale (Hauck et al., 1990) and the Wake Forest Physician Trust Scale (Hall et al., 2002). Both have yielded valid and reliable data. Patient questionnaires have advantages: they obtain information about physicians’ behaviors in actual practice, they seem to be acceptable to physicians, and they may prompt changes in behaviors assessed. As part of the American Board of Internal Medicine Continuous Professional Development Program, patients assessed physicians’ professionalism and communication skills (Lipner et al., 2002). Of physician participants, 61% thought the module had provided them with a valuable learning experience, 82% said they would continue to seek feedback from patients and peers, and 42% reported they would change their professionalism and communication strategies. Since the study participants were volunteers, the findings may not be representative of all board certified internists.

Practice points

- At least 88 professionalism assessments have been used in medical education since 1982.
- Professionalism assessments may be organized into content area addressed (i.e. ethics, personal characteristics, comprehensive professionalism and diversity) and type of outcome examined (i.e. affective, cognitive, behavioral and environmental).
- Useful approaches to assessing medical student professionalism include a longitudinal, performance-based method and behavioral simulations with post-encounter probes.
- Useful approaches to assessing resident physician professionalism include a 360-degree assessment and a cognitive assessment of professionalism.
- Useful approaches to assessing practicing physician professionalism include patient questionnaires and a cognitive assessment of professionalism.

Notes on contributors

DIERDRE C. LYNCH, RhD, is a Research and Evaluation Specialist at the Accreditation Council for Graduate Medical Education (ACGME) in Chicago. Her research interests include evaluating health professions education and curriculum improvement.

PATRICIA M. SURDYK, PhD, is Senior Project Manager at the ACGME in Chicago. She is leading the ACGME’s efforts to study how professionalism can be fostered and assessed in graduate medical education.

ARNOLD R. EISER, MD FACP, is Vice President of Medical Education at Drexel University College of Medicine, Philadelphia. His research interests include developing and assessing ethics and professionalism curricula.

References


