Measuring Professionalism: A Review of Studies with Instruments Reported in the Literature between 1982 and 2002

J. Jon Veloski, MS, Sylvia K. Fields, EdD, RN, James R. Boex, PhD, and Linda L. Blank

Abstract

Purpose
To describe the measurement properties of instruments reported in the literature that faculty might use to measure professionalism in medical students and residents.

Method
The authors reviewed studies published between 1982 and 2002 that had been located using Medline and four other databases. A national panel of 12 experts in measurement and research in medical education extracted data from research reports using a structured critique form.

Results
A total of 134 empirical studies related to the concept of professionalism were identified. The content of 114 involved specific elements of professionalism, such as ethics, humanism, and multiculturalism, or associated phenomena in the educational environment such as abuse and cheating. Few studies addressed professionalism as a comprehensive construct (11 studies) or as a distinct facet of clinical competence (nine studies). The purpose of 109 studies was research or program evaluation, rather than summative or formative assessment. Sixty five used self-administered instruments with no independent observation of the participants' professional behavior. Evidence of reliability was reported in 62 studies. Although content validity was reported in 86 studies, only 34 provided strong evidence. Evidence of concurrent or predictive validity was provided in 43 and 16 studies, respectively.

Conclusions
There are few well-documented studies of instruments that can be used to measure professionalism in formative or summative evaluation. When evaluating the tools described in published research it is essential for faculty to look critically for evidence related to the three fundamental measurement properties of content validity, reliability, and practicality.

Acad Med. 2005; 80:366–370.

Professionalism and related personal attributes, such as ethics, humanism and personal values have played a central role in the major critiques and calls for reform in medical education over the past century.[1] Although medical schools and residency programs have always striven to foster the professional growth of young physicians, the economic and social forces influencing health care in recent decades have focused renewed attention on the importance of professionalism. The lists of "competencies" currently recommended for the curricula of medical schools by the Association of American Medical Colleges and for residency programs by the Accreditation Council for Graduate Medical Education embrace professionalism as one distinct facet of physician competence.[2,3] Individual specialty groups, such as the American Board of Internal Medicine, have had long-standing commitments to professionalism.[4–7]

Substantive discussions about the development of professionalism in medical students and graduate medical education inevitably lead to questions about measurement and evaluation.[8] Whether the goal is to evaluate these individuals as part of a formal education program or to provide information for self-assessment, counseling, or remediation, faculty in medical schools and graduate medical education programs seek credible instruments. The minimum evidence required to support such instruments in medical education includes content validity, as judged by national experts, and high reliability with acceptable levels of measurement error.[9] Additional evidence of empirical validity, such as concurrent, predictive, or construct validity, provides the most complete assurance of an instrument's quality.

Attempts to develop instruments to measure professionalism exist. Arnold[10] cites over 170 articles in summarizing the state of the art in 2002. She affirmed that a concept of professionalism had been described in the literature and is available for the development of assessment tools. Although she refers to a rich array of existing assessment tools, she emphasizes the need to strengthen their measurement properties. She concludes by reminding readers that it would be impossible to answer questions about the efficacy of educational efforts related to professionalism without solid instruments.

We undertook this review to analyze the measurement goals and the reliability and validity of the instruments used in studies related to the measurement of professionalism reported in the literature over the past two decades.

Mr. Veloski is director of medical education research at the Center for Research in Medical Education and Health Care, Jefferson Medical College, Philadelphia, Pennsylvania.

Dr. Fields is the executive director of the Savannah Health Mission, Savannah, Georgia, and formerly a senior research associate and director of community programs at Jefferson Medical College, Philadelphia, Pennsylvania.

Dr. Boex is the director of the Office of Health Services at Northeastern Ohio Universities College of Medicine, Rootstown, Ohio.

Ms. Blank is the Senior Vice President of the American Board of Internal Medicine Foundation, Philadelphia, Pennsylvania.

Correspondence should be addressed to Mr. Veloski at the Center for Research in Medical Education and Health Care, Jefferson Medical College, 1025 Walnut Street, Suite 119, Philadelphia, PA 19107; e-mail: ⟨jon.veloski@jefferson.edu⟩.

Method

The primary source of data for the review consisted of articles identified in another published review.[11] In that review, the investigators had searched Medline, ERIC, HAPI, PsychINFO, and TIMELIT for studies published between 1982 and 2002 using 28 search terms, including "professionalism," "duty," or "ethics" in combination with "assessment," "evaluation," or "measurement." The reference lists of relevant articles also had been searched manually.

Lynch and colleagues[11] used two criteria for selection: the study must have involved medical students, housestaff, medical schools, or teaching hospitals; and the study must have provided empirical evidence based on the use of an instrument that included at least two items or a defined set of qualitative categories. They excluded studies that appeared to address only communication skills and those describing highly specific issues in professionalism, such as physicians' responses to "do-not-resuscitate" orders.

To build on their findings, we manually searched the contents of *Academic Medicine* in late 2002 and early 2003 to locate additional relevant studies.

Data extraction

We developed a four-page data extraction form using nine forced-choice and two open-ended items. The first item, which addressed the definition of professionalism, offered three options: professionalism as a comprehensive construct, professionalism as one facet of clinical competence, or professionalism as an array of separate elements. If the reviewer checked the third option, they were instructed to list key terms to describe the elements of professionalism measured in the study. Arnold[10] developed this approach to classifying studies of professionalism among three broad types in her review.

Three items, which were related to the goals of measurement, addressed the target of assessment (i.e., who or what the study measured), the respondent group providing the data (i.e., who filled out the questionnaires), and the primary purpose of the assessment (i.e., formative, summative, research, or program evaluation). The next four items addressed evidence related to reliability and validity reported in support of the instrument, and the reviewer's judgment of the quality of the validity evidence. A final set of items collected the reviewer's ratings of the instrument's practicality and implications for future research. "Practicality" was defined as ease of administration; cost-effectiveness; and acceptance by participants, observers, and academic leaders. The draft form and instructions were pretested in two iterations and revised accordingly based on the comments of individuals at Jefferson Medical College, the National Board of Medical Examiners, and the American Board of Internal Medicine Foundation.

We identified a panel of 12 highly qualified reviewers who possessed a formal background in measurement and research as well as a record of professional activity and peer-reviewed publications in medical education. During the first phase of the review, each member of the panel completed data extraction forms for approximately ten studies. We calculated an overall rating for each study by summing their ratings for validity evidence, practicality, and implications for future research. We selected the studies rated in the top and bottom quintiles of this overall rating, as well as a random sample of 25 studies in the middle three quintiles, and reassigned them to another reviewer for a second, independent review.

One of the authors (SKF) and a research assistant under her supervision prepared a structured, one-page summary of each article. Each summary included a description of the instrument, the characteristics of the sample, the methods of administration, any scoring procedures, and a synopsis of key evidence related to the instrument's reliability and validity. Finally, the reviewers' overall ratings of the practicality and research implications of each instrument were added. Copies of the data extraction form and a set of the 134 summaries are available either from the authors or online at ⟨http://www.abimfoundation.org/pdf/MPP_Summaries.pdf⟩.

Data synthesis

One of the authors (JJV) compared the completed forms against the structured summaries to affirm congruence between the reviewers' responses on the data extraction forms and the narrative summaries. Inconsistencies were resolved by referring to the published article. Similarly, the consistency in classification of the definition of professionalism, measurement target, source of data, reliability estimates, and validity evidence was checked. Disagreements were resolved by referring to the article.

The forms were entered into a computer spreadsheet. Frequency distributions and cross tabulations were prepared using Stata software (version 8.0).

Interrater reliability of the reviewers' coding

There was very high (> 90%) agreement between the reviewers' responses for the target, source, and purpose of the publications with the narrative summaries, and between the two reviews of the same article. There was also very high (> 90%) agreement in the general classification of the definition of professionalism. However, the identification of keywords related to the elements of professionalism was less consistent. As noted in Table 1, inconsistencies in the responses of multiple reviewers were resolved by one of the authors (JJV) based on the title of the published article or stated purpose of the study.

There was less consistency in the reviewers' responses for reliability. Disagreements about reliability often involved either incomplete reporting in the article or citations to previous publications without reporting specific values. The results reflect the responses of the more lenient of the multiple reviewers, those willing to infer reliability even if the report of evidence in the article was vague, incomplete, or cited without detail.

There was even less consistency in the responses for validity. Here, the reviewers were instructed to locate evidence of content, construct, and criterion-related validity and to judge whether this evidence met the published standards promulgated by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.[12] The results suggested that the reviewers applied differing criteria for the presence of validity, especially content validity. The responses on the form and the informal reports of the reviewers themselves indicated that the published standards left a great deal of room for interpretation. Again, the results reported here reflect the responses of the more lenient of multiple reviewers,

Table 1

Elements of Content Definition Used to Develop Instruments in 134 Studies Related to the Measurement of Professionalism, 1982–2002

Definition	No. of studies
Specific attributes of a professional*	
Ethics, decision making moral reasoning†	48
Humanism	11
Multiculturalism	8
Empathy	4
Values	4
Deception in patient relationships, attitudes toward	3
Indigent, care for	2
Trust	2
Attitudes and communication	1
Confidentiality of patient data	1
Contact with patients, appropriate/inappropriate	1
Emotional intelligence	1
Mental health	1
Self-assessment	1
Mixed attributes	6
Professionalism as one facet of competence	9
Professionalism as a comprehensive construct	11
Other phenomena	
Abuse and harassment of students, housestaff	7
Patient satisfaction	5
Cheating	4
Uncertainty, attitudes toward	2
Cynicism	1
Turfing	1
Total	134

* The list of attributes of a professional is empirical, having been derived from titles of the article, abstracts, purpose, or key terms provided by the reviewers. Each article was assigned to one category. However, the attributes in this list are not necessarily mutually exclusive, and the list is not intended to be exhaustive.
† The heterogeneous category "ethics" includes ethics, ethical decision-making skills, moral reasoning abilities, and related attributes.

that is, those of reviewers willing to accept validity evidence even if the report of evidence in the article was vague, incomplete, or cited without detail.

In two items the reviewers rated their overall impressions of the practicality and future implications of the concept or model presented in the study. There was no consistent agreement in the reviewers' ratings of the practicality and implications for future research and development of individual articles.

Results

A total of 134 studies were identified that reported empirical findings based on measurements of medical professionalism including specific attributes of professionals and related phenomena. More than half (72) were published after 1995, with about one-third (46) published in 2000 or later. About one-fifth (23) had been published before 1990. The most frequently referenced journals were *Academic Medicine* (44) and its predecessor, the *Journal of Medical Education* (8).

Definitions of professionalism

Over three-fourths (94) of the studies focused on specific elements of professionalism. More than one-third (48) involved ethics, ethical decision making, and moral reasoning (see Table 1). Eleven studies involved the measurement of humanism, while the remaining studies were distributed across a broad spectrum of distinct elements. Only 11 studies addressed the measurement of professionalism as a comprehensive construct. In these studies, the respondent was either provided with representative examples of professional or unprofessional behavior or was allowed to decide what specific behaviors would be considered subject to a broad definition of professionalism. Finally, nine studies measured professionalism as one distinct facet of clinical competence as distinguished within a set of competencies, including, for example, knowledge, clinical skills, and communications skills.

A group of 20 studies involving other phenomena often associated with professionalism is identified separately in Table 1. Several reviewers recommended that these studies be differentiated in the analysis.

Target of assessment

The target of assessment varied widely as summarized in the columns of Table 2. The largest subset of studies (63) was directed toward the measurement of groups, either of students (27), housestaff (14), physicians (1), or combinations (21). Another 25 studies involved measurements of the environments of medical schools or teaching hospitals. Inspection of the summaries of the studies confirmed that most studies of groups and environments involved anonymous opinion surveys. Only 46 of the 134 studies involved measurements of individuals using instruments such as rating forms that yielded a score or vector of scores for each participant.

Source of data

The rows of Table 2 distinguish the source of data from the target of assessment. In about half of the studies (65), the participants provided their own data through self-report instruments, such as knowledge tests, personality tests, inventories of personal experiences, attitude/opinion surveys, or self-assessments of competence. The most frequent types of participants were medical students (36), followed by housestaff (10) and physicians (3). The remaining studies involved combinations of these three groups.

Sixty nine studies used data collected by independent observers such as faculty (16), medical students (14), or patients (9). The remaining 30 of these 69 studies used a wide array of other independent

Table 2

Target of Assessment and Source of Data in 134 Studies on the Measurement of Professionalism, 1982–2002

	Target of assessment								
	Individual			Group					
Source of data	Student	Housestaff	Physician	Students	Housestaff	Physicians	Combinations	Environment	Total
Self									
(Self-assessments, self-reports, knowledge or personality tests)	14	3	3	22	7		16		65
Independent observers									
Medical students				1				13	14
Physicians, faculty	7	4		2	3				16
Patients	1		3		2		3		9
Other (house staff, nurses, standardized patients or combinations)	7	3	1	2	2	1	2	12	30
Total	29	10	7	27	14	1	21	25	134

observers including housestaff, nurses, standardized patients, or combinations of different types of observers. Examples of the instrumentation used by observers included global rating forms, observational rating forms, observational checklists, focus-group protocols, interview guides, and reports of exceptional behavior.

Primary purpose of assessment

Research or program evaluation was the primary purpose of the vast majority (109) of studies. Only a handful (14) were directed toward summative evaluation, and even fewer (11) stated a primary goal of formative evaluation. The end product of most studies was aggregate statistical data—rather than individual scores—to be used for research or program evaluation.

Reliability and validity

As summarized in Table 3 approximately half (62) of the studies reported estimates of reliability, including internal consistency, generalizability, interrater reliability or test-retest reliability. However, 72 reported no information about reliability or any attempt to estimate errors of measurement.

The reviewers found some evidence of content validity in the majority (86) of studies, and 34 of these provided strong evidence of content validation with a broad sample of experts that approached national standards. However, 48 included no attention to content validity, which is the foundation of validation in mental measurements. Although almost half (61) of the studies provided some evidence of construct validity, only about one-third (43) reported on concurrent validity, and a handful (16) considered predictive validity. Overall, using a five-point scale ranging from very low to very high, the reviewers rated the strength of validity evidence as high or very high for only 15 of the 134 studies.

Practicality

We defined practicality as ease of administration; cost-effectiveness; and acceptance by participants, observers, and academic leaders. The reviewers reported that about one-quarter (32) of studies provided strong evidence of practicality, as supported by, for example, operational use at a medical school or residency program or use at multiple sites. Some evidence of practicality was reported in about two-thirds of studies.

Discussion

Our search of the literature spanning more than two decades located 134 articles that reported the results of empirical studies designed to measure professionalism in medicine. A review panel of qualified experts in medical education research used a formal protocol to extract data on the instruments described in the studies, including the definition of professionalism; purpose of assessment; and evidence of reliability, validity, and practicality. The vast majority of the instruments were designed or adapted for research or program evaluation. Many tools were designed to measure learning environments or groups of students and physicians rather than individuals. There was limited attention to the estimation of reliability and errors of measurement. Most often, validation of the content measured by the instruments rested solely on the judgment of convenience samples of local experts, with limited attention to construct, concurrent, or predictive validity. Few studies provided evidence of practicality beyond a single trial at a single site.

On one hand, some readers will find these results surprising, even disappointing. Discussions of the assessment of professionalism sometimes imply the need to gather data on individuals that can be used to provide feedback; to guide referrals to remedial programs; or to inform decision-makers on grading, academic promotion, licensing, or certification decisions. Few instruments met the minimal criteria of content validity, reliability, and practicality that would support their operational use for academic decision making.

On the other hand, these results are not unanticipated. The primary purpose of

Table 3

Frequency and Types of Reliability and Validity Evidence Reported in 134 Studies of Professionalism, 1982–2002*

Study type	No. of studies
Reliability	
Internal consistency	33
Inter-rater	24
Test–retest	16
Other	11
None	72
Validity	
Content	86
Construct	61
Concurrent	43
Predictive	16
None	19

* Other approaches to reliability estimation include generalizability studies, other analysis of variance studies and citations to reliability studies without specifications. Frequencies of each type of reliability and validity do not sum to 134 because multiple types were reported within some studies.

the vast majority of the studies reviewed was research or program evaluation. They were designed either to explore one specific aspect of professionalism or to implement some instructional activity such as a course to enhance professionalism. Measurement and evaluation were secondary to the goal of most studies. Correspondingly, it is important to emphasize that the findings of this review do not reflect on the overall quality of the research reported in the studies. The review reported here was designed to analyze the measurement properties of the instruments and other methods used to measure professionalism and to judge these properties in relation to accepted standards.

There are several limitations of this review that should be addressed in future analyses of the array of instruments available to measure professionalism. First, the set of 134 studies is a function of the scope and timing of the literature search. We believe this set is comprehensive and accurately represents the universe of published studies in this area during this time period. The summaries of the articles, which are available at ⟨http://www.abimfoundation.org/pdf/MPP_Summaries.pdf⟩, provide evidence of the broad representation of the studies and instruments covered by this review. Nevertheless, the complexity of the construct of professionalism in medicine and the complexity of the English language invite different interpretations that may imply additional keywords in future searches. Furthermore, the fact that over one-third of the studies were published after 2000 implies that additional papers had been published after the search was completed and that significant research is in progress and is yet to be published. Second, our review concentrated on professionalism in medical students, housestaff, and physicians. It focused on the literature of medical education. It is possible that studies of these instruments or other instruments in other health professions or even other professionals exist that may provide important information to medical educators. Finally, the unit of analysis was studies rather than instruments. The most comprehensive review would include reports of all studies involving each instrument, not only those studies related to the measurement of professionalism in medicine.

The findings of this review have implications for medical schools and residency programs that are looking to the literature for new or proven methods of measuring professionalism for use in their educational programs. *Caveat emptor.* When evaluating the tools described in published research it is essential to look critically for evidence related to the three fundamental properties of content validity, reliability, and practicality. Content validity must be demonstrated by a systematic analysis of the domain being measured, involving a representative cross-section of content experts. There must be evidence that the instrument adequately samples the content of this domain. Reliability estimates must be explicitly reported with sufficient information about the variation in scores to demonstrate that errors of measurement are within acceptable limits. Finally, evidence of practicality including cost, ease of administration, and acceptance by trainees and faculty must be provided based on field tests in a representative sample of subjects and settings.

This study was supported in part by the American Board of Internal Medicine Foundation. The authors thank the reviewers: Mark A. Albanese, PhD; Louise Arnold, PhD; Barbara Barzansky, PhD; Jan Carline, PhD; Carol Elam, EdD; Shiphra Ginsburg, MD, MEd; Larry Gruppen, PhD; John Littlefield, PhD; Deirdre Lynch, PhD; William C. McGaghie, PhD; David Stern, MD, PhD; and Reed Williams, PhD. The authors also thank Stephen Clyman, MD, Robert Galbraith, MD, who assisted with the planning of the review process, and, Emily Gavin, Caryl Johnston, MEd, Gail Andrus, Mary R. Robeson, MS, and Sandra K. Maxwell, who helped at various stages with data handling and manuscript preparation.

References

1 Christakis NA. The similarity and frequency of proposals to reform US medical education: constant concerns. JAMA. 1995;274:706–11.

2 Medical School Objectives Project. Learning objectives for medical student education. Guidelines for medical schools: Report I of the Medical School Objectives Project. Acad Med. 1999;74:13-8.

3 Leach D. Competence is a habit. JAMA. 2002; 287:243–4.

4 Stobo JB, Blank LL. Project professionalism: staying ahead of the wave. Am J Med. 1994; 97(6):i–iii.

5 Arnold EL, Blank LL, Race KE, Cipparrone N. Can professionalism be measured? The development of a scale for use in the medical environment. Acad Med. 1998;73:1119–21.

6 Medical professionalism in the new millennium: a physician charter. Ann Intern Med. 2002;136:243–6.

7 Medical professionalism in the new millennium: a physician charter. Lancet. 2002;359: 520–2.

8 Whitcomb ME. Fostering and evaluating professionalism in medical education. Acad Med. 2003;77:473–4.

9 Hubbard JP. Measuring Medical Education. The Tests and the Experience of the National Board of Medical Examiners. 2nd ed. Philadelphia: Henry Kimpton Publishers, 1978.

10 Arnold L. Assessing professional behavior: yesterday, today, and tomorrow. Acad Med. 2002;77:502–15.

11 Lynch D, Surdyk P, Eiser A. Assessing professionalism: a review of the literature. Med Teach. 2004;26:366–73.

12 American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association, 2002.