Accuracy of Physician Self-assessment Compared With Observed Measures of Competence

A Systematic Review

David A. Davis, MD
Paul E. Mazmanian, PhD
Michael Fordis, MD
R. Van Harrison, PhD
Kevin E. Thorpe, MMath
Laure Perrier, MEd, MLIS

Context Core physician activities of lifelong learning, continuing medical education credit, relicensure, specialty recertification, and clinical competence are linked to the abilities of physicians to assess their own learning needs and choose educational activities that meet these needs.

Objective To determine how accurately physicians self-assess compared with external observations of their competence.

Data Sources The electronic databases MEDLINE (1966-July 2006), EMBASE (1980-July 2006), CINAHL (1982-July 2006), PsycINFO (1967-July 2006), the Research and Development Resource Base in CME (1978-July 2006), and proprietary search engines were searched using terms related to self-directed learning, self-assessment, and self-reflection.

Study Selection Studies were included if they compared physicians' self-rated assessments with external observations, used quantifiable and replicable measures, included a study population of at least 50% practicing physicians, residents, or similar health professionals, and were conducted in the United Kingdom, Canada, United States, Australia, or New Zealand. Studies were excluded if they were comparisons of self-reports, studies of medical students, assessed physician beliefs about patient status, described the development of self-assessment measures, or were self-assessment programs of specialty societies. Studies conducted in the context of an educational or quality improvement intervention were included only if comparative data were obtained before the intervention.

Data Extraction Study population, content area and self-assessment domain of the study, methods used to measure the self-assessment of study participants and those used to measure their competence or performance, existence and use of statistical tests, study outcomes, and explanatory comparative data were extracted.

Data Synthesis The search yielded 725 articles, of which 17 met all inclusion criteria. The studies included a wide range of domains, comparisons, measures, and methodological rigor. Of the 20 comparisons between self- and external assessment, 13 demonstrated little, no, or an inverse relationship and 7 demonstrated positive associations. A number of studies found the worst accuracy in self-assessment among physicians who were the least skilled and those who were the most confident. These results are consistent with those found in other professions.

Conclusions While suboptimal in quality, the preponderance of evidence suggests that physicians have a limited ability to accurately self-assess. The processes currently used to undertake professional development and evaluate competence may need to focus more on external assessment.

JAMA. 2006;296:1094-1102 www.jama.com

SELF-ASSESSMENT AND SELF-directed, lifelong learning have long been mainstays of the medical profession—they are activities presumed to be linked closely to the quality of care provided to patients.[1] Physicians in the United States must demonstrate their engagement in lifelong learning by choosing and participating in continuing medical education (CME) activities[2] and acquiring CME credit, which is mandated by the majority of state medical boards under the rubric of states' medical practice acts.[3] The American Medical Association's Physicians Recognition Award certificate,[4] which is based on CME participation, meets the CME requirements of the Joint Commission on Accreditation of Healthcare Organizations related to hospital accreditation.

Self-assessment and lifelong learning were adopted by the American Board of Medical Specialties explicitly as 1 of 4 elements in its Maintenance of Certification program.[5] Furthermore, diplomates of the American Board of Internal Medicine who choose to recertify

For editorial comment see p 1137.

CME available online at www.jama.com

Author Affiliations: Knowledge Translation Program of the Li Ka Shing Knowledge Institute at St Michael's Hospital (Dr Davis and Mr Thorpe), Departments of Health Policy, Management, and Evaluation (Dr Davis), Family and Community Medicine (Dr Davis), and Public Health Sciences (Mr Thorpe), and the Office of Continuing Education and Professional Development (Ms Perrier), University of Toronto, Toronto, Ontario; Departments of Family Medicine and Epidemiology and Community Health, School of Medicine, Virginia Commonwealth University, Richmond (Dr Mazmanian); Center for Collaborative and Interactive Technologies, Baylor College of Medicine, Houston, Tex (Dr Fordis); and Department of Medical Education, University of Michigan, Ann Arbor (Dr Harrison).
Corresponding Author: Laure Perrier, MEd, MLIS, University of Toronto, 500 University Ave, 6th Floor, Toronto, Ontario, Canada M5G 1V7 (l.perrier@utoronto.ca).

must complete 10-year cycles of recertification, a process focused on continuous professional development that requires the capacity of physicians to self-assess.[6] In graduate medical education, the issue of practice-based learning and improvement based on self-assessment is a central tenet of professional development in Canada,[7] the United States,[8] and in other countries.[9]

Each of the elements in this chain—the emphasis on self-assessment, self-directed lifelong learning, the acquisition of CME credits and their use for medical relicensure, accreditation, and ongoing certification—is heavily dependent on the ability of physicians to determine their own learning needs and find resources to meet them. However, a recent review of theory-oriented literature of self-assessment (including the health professions) raises serious questions about the failure of professionals to generate summary judgments of their performance in any regular or consistent fashion, a critical requirement for a self-regulating profession.[10]

While the term self-assessment is used to describe many types of activities, we were interested in considering the aspects of "self-rating" or "self-audit" in contrast to the use of self-administered examination of knowledge or clinical performance. To our knowledge, no systematic reviews of studies of this type of physician self-assessment compared with external observation as a reference standard exist. We, therefore, reviewed the literature to determine how accurately physicians self-assess compared with external observations of their competence.

METHODS

Data Sources

The databases of MEDLINE (1966-July 2006), EMBASE (1980-July 2006), CINAHL (1982-July 2006), PsycINFO (1967-July 2006), the Research and Development Resource Base in CME (1978-July 2006),[11] and proprietary search engines were searched using the terms *self-directed learning*, *self-assessment*, *self-reflection*, *self-rating*, *reflective practice*, *multisource feedback*, and related terms. Indexing of this topic in the literature databases is limited and few relevant subject headings were available, thus searching relied on identifying key words through seminal articles and expert consensus. Hand-searching was performed by reviewing references of retrieved articles. The complete search strategies are available on request.

Data Selection

Studies that focused on a comparison between physicians' self-assessments as determined by self-ratings and 1 or more external measures of related competencies were included. Studies were selected that used such self-assessments (ie, physician perceptions or predictions of knowledge, skill, or performance) compared with an external, well-described measure (eg, objective structured clinical examinations [OSCEs] in the same domain), or compared observed performance ratings. In addition, included studies had to: use quantifiable and replicable measures; have a study population of at least 50% practicing physicians, residents, or similar health professionals such as nurse practitioners and physician assistants to be able to generalize to these groups; and be conducted in the United Kingdom, Canada, United States, Australia, or New Zealand, which have similar training requirements, maintenance of competence programs, languages, and CME requirements. We excluded studies that were comparisons of self-reports; evaluated medical students or primarily examined other health professionals; described the development or testing of self-assessment measures; assessed physician beliefs about patient status; or were self-assessment programs of specialty societies that provide feedback to physicians on tests of competence. Studies conducted in the context of an educational or quality improvement intervention in the subject area of the assessment were included only if comparative data were obtained before the intervention.

Data Extraction

The following information was extracted from each article: study population; content area and self-assessment domain of the study; methods used to measure the self-assessment of study participants; methods used to observe or measure participants' competence or performance; existence and use of quantifiable measures; and study outcomes. When available, data within studies that might explain the association between the self- and external assessments were sought. None of the data extraction was performed in a blinded fashion. A meta-analysis was not performed because the conceptual constructs within the domains of self-assessment were varied and assessed different skills using varying measures.

The methodological quality of the articles was assessed by determining (1) whether there was sufficient description to permit replication of the study population, (2) the content domain of the self-assessment, and (3) the explicit indication of blinding of the external observation to the self-assessment. In addition, the sampling frame on which the study population was based, the methods used in the self- and external assessments, the identification of pilot testing or use of previously validated methods by the authors of the studies, and the presence and appropriateness of statistical tests were determined.

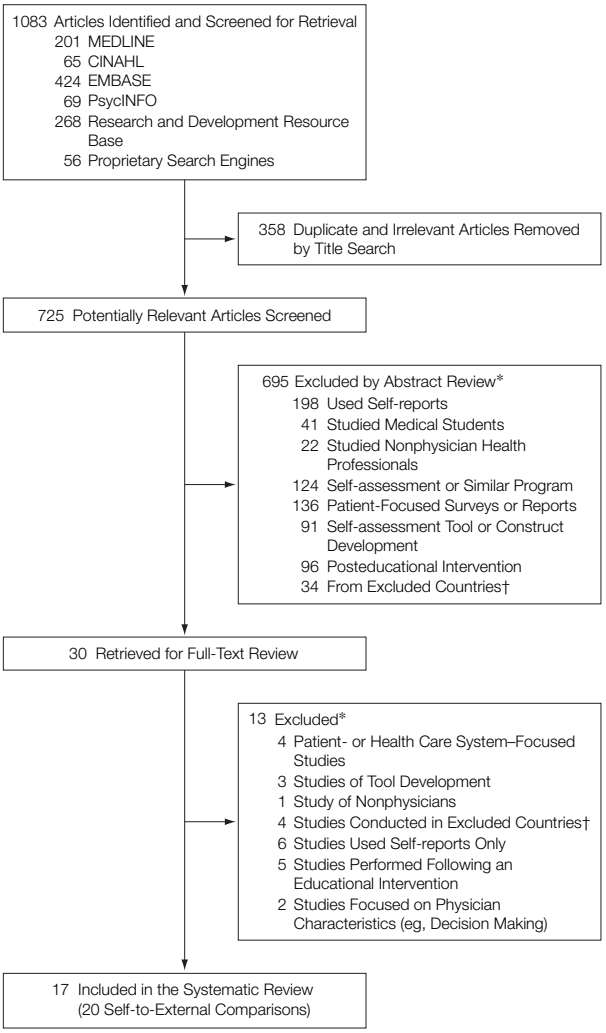
The literature search was performed by one of the authors (L.P.) and duplicated by an independent information specialist. One of the authors (D.A.D.) determined the inclusion criteria. The inclusion criteria were applied to the abstracts of all articles by 2 of the authors (D.A.D., P.E.M.) and corroborated by another author (L.P.). Data extraction methods were developed by one of the authors (D.A.D.) and were applied by 2 of the authors (D.A.D., P.E.M.) to those articles that met the inclusion criteria. Disagreements about search criteria, data extraction, and classification of study results were resolved by consensus.

RESULTS

Search Results and Article Overview

The search strategies yielded 725 articles after removal of duplicate and irrelevant studies by title search (FIGURE). After applying inclusion criteria to the abstracts of these articles and excluding studies that were based on self-reports, evaluated medical students or other health professionals who did not fit the inclusion criteria, were patient-focused, or were reports on the development of self-assessment tools and constructs, posteducation assessments, or self-assessments of specialty society programs, there were 30 articles published between 1988 and 2005. After review of the full text of these articles, another 13 were excluded based on these criteria or because they focused on physician characteristics such as the decision-making process or were conducted in excluded countries, leaving 17 articles[12-28] that met all of the inclusion criteria (Figure). Three studies used 2 external comparisons each,[12,15,23] resulting in 20 comparisons between self- and external assessment.

Figure. Search and Selection of Included Studies



*Some articles excluded for multiple reasons.
†Countries other than the United Kingdom, Canada, United States, Australia, and New Zealand.

The majority of the studies reported findings related to clinical medicine, such as procedural skills, palliative care, and general medical knowledge.[12-17,21-25,27,28] One study focused on teaching abilities,[18] another focused on cultural competencies,[19] and 2 studies focused on evidence-based medicine[20,26] (TABLE 1). Six studies[13,18-20,24,27] examined the self-assessment abilities of practicing physicians and one study[16] focused on practicing physicians, physician assistants, and nurse practitioners. The study by Leopold et al[16] was conducted in the context of a specific, focused educational intervention. The remainder of the studies focused on graduate physician-trainees[12,14,15,17,21-23,26] or both trainees and practicing physicians[25,28] (TABLE 2 and TABLE 3).

Domains of Self-assessment

The 20 comparisons between self- and external assessment were divided into 3 constructs of self-assessment. Six studies[14,15,17,20,26,28] focused on *predictive self-assessment*, which is the ability of the physician to predict his/her performance on a future competency-based assessment.[12,16,18,19,21-25]

Nine comparisons were in the construct of *summative and retrospective self-assessment*. In 2 studies,[12,16] participants were asked to rate their performance in a recently completed simulation exercise, which was later compared with the ratings given by external observers. Seven studies asked participants to provide mental representations of themselves over time (ie, competent in clinical medicine,[21-25] competent as surgical teachers,[18] and cultural-linguistic competence[19]) compared with the perceptions of performance in these areas by residents,[18] patients,[19] faculty supervisors and staff members,[22,23] objective tests,[21,23,24] or chart audit.[25]

The final construct was *concurrent self-assessment*. Two studies[13,27] were in this category, asking physicians to self-identify current learning needs. In each, the process involved included a reflection on performance, knowledge, or skills in familiar situations.

Methods of Self-assessment

Self-ratings were conducted in the studies by questionnaire, checklists, or survey and focused on learning needs,[13,27] confidence in performing procedures,[14-16] general clinical skills,[12,21-25,27,28] medical and critical appraisal knowledge,[17,20,26] and nonclinical competencies (eg, teaching skills[18] and cultural-linguistic competencies[19]).

Methods of External Assessment

Studies compared physician self-ratings with stable external objective measures such as OSCEs,[14,15] standardized patients,[12,21] simulations,[16] performance on in-training or other examinations,[17,23,24,26-28] chart audit,[25] or the ability to explain concepts of evidence-based medicine to a blinded interviewer.[20] Studies also compared physician self-ratings to data derived from structured interviews aided by physician diaries[13] or ratings by stakeholders including residents,[18] patients,[19] or faculty supervisors.[12,15,22,23] In 3 of the studies,[12,18,23] the instruments used in the physician self-ratings closely matched the instruments used for comparative purposes.

Methods of Comparing Self- and External Assessments

The studies demonstrated heterogeneity in their choice of comparisons and use of statistical methods. The studies reported either use of descriptive statistics[13,20,21,24] or inferential statistics.[12,14-19,22,23,25-28] Two studies[12,22] did not identify the statistical tests applied. One study[12] used the same instrument for self- and external measurements, thus permitting precise tests of agreement.

Accuracy of Self-assessment

Of the 20 comparisons between self- and external assessment, 13 demonstrated little, no, or an inverse relationship between self-assessment measures and other indicators.[12-18,20,21,23,26,27] Six studies used ratings of global confidence compared with performance of procedures, using behavioral checklists for rating performance in the assessment of dementia,[12] procedural skills,[14-16] health promotion counseling,[21] and critical care skills.[23] Two studies used structured interviews as an external measure, identifying a lack of congruence between self-assessment and external observation in detecting learning needs in palliative care physicians[13] and in physicians' abilities to explain evidence-based medicine terms.[20] Three studies used tests such as multiple choice examinations in critical care,[23] standardized articles to test critical appraisal skills,[26] and true-false tests in 3 areas of primary care.[27] One study demonstrated a failure of surgeons to judge the perceptions of residents relative to the surgeons' teaching skills.[18] Overall, the proportion of studies reporting little, no, or inverse relationships did not appear to vary by

Table 1. Studies Examining the Self-assessment of Teaching Abilities, Cultural Competencies, and Evidence-Based Medicine

Source	Population; Participation Rate	Content Area; Domain of Self-assessment; Type of Self-assessment*	Self- and External Assessment Methods	Deficiencies in Methodological Quality; Study Outcomes†
Claridge et al,[18] 2003	23 US attending surgeons; 78%	Surgical teaching skills; Ability to self-assess past teaching performance in operating rooms, surgical wards, and other settings; Summative	Staff surgeon questionnaire using 5-point Likert rating scale to measure teaching skills; Resident questionnaire using 5-point Likert rating scale to measure teaching skills of supervising surgeons	Nonparticipants not described, no pilot testing of self- or external assessment, paired *t* tests not indicated or inappropriately used; Significant differences in self-ratings of 11 of 17 attending surgeons vs residents' ratings, attending surgeons who participated in self-evaluation were significantly more highly rated than those who did not
Fernandez et al,[19] 2004	48 US family physicians or general internists; 100%	Cultural and linguistic competencies; Mental representation of self as having language skills, cultural competency, and effectiveness with culturally diverse patients; Summative	Physician questionnaire to measure cultural and linguistic competencies using 5-point Likert rating scale; Patient satisfaction using an established interpersonal process of care instrument	No pilot testing of self-assessment method, calculation of OR unclear; Positive association between physician self-perception of abilities and patient satisfaction (adjusted OR, 5.25; 95% CI, 1.59-17.27)
Stern et al,[26] 1995	62 US internal medicine residents; 52%	Critical appraisal; Performance on a test of critical appraisal; Predictive	Resident self-assessed competence using Likert rating scale; Performance measured by critical appraisal of sample journal article previously reviewed by experts	Nonparticipants not described; unclear description and no pilot testing of self-assessment questionnaire, no *P* values; No significant correlation between self-assessments and actual scores ($r = 0.15$)
Young et al,[20] 2002	50 Australian general practitioners; NA	Concepts of evidence-based medicine; Performance in explaining terms used in evidence-based medicine; Predictive	Physician self-rating of abilities; Explanation of terms to a blinded interviewer using a checklist	Total and sample population not described, no pilot testing of assessment measures; Verbal explanations rarely matched (and never exceeded) general practioners' self-rated abilities to explain terms

Abbreviations: CI, confidence interval; NA, not available; OR, odds ratio.

*The 3 discrete types of self-assessment used in this review were predictive (the ability of the physician to predict his/her performance on a future competency-based assessment), summative (retrospective evaluation of performance or abilities), and concurrent (identification of current learning needs).

†Unless otherwise indicated, all qualitative and descriptive measures were sufficiently described to permit replication. Most studies did not blind the external assessment to the self-rating (the study by Young et al[20] did blind the external assessment).

level of training or experience or by year of study.

In contrast, 7 comparisons[12,15,19,22,24,25,28] demonstrated positive associations between self-assessment and most external observations. Three found consistency (and little variability) between postperformance self-ratings by residents and observer ratings on global or general levels of performance in assessing dementia,[12] basic clinical skills,[15] and in competency in managing the psychological aspects of family practice.[22] Two studies demonstrated positive associations between self-rated expertise and the diagnosis of childhood sexual abuse by physical exmination,[24] or "comfort" in recognizing the diagnostic features of smallpox.[28] Both studies used case vignettes and photographs of related physical findings. One study[19] displayed a strong association between physician self-rating of language and cultural competency and patient reports of the interpersonal process of care using a standardized instrument. Another study[25] demonstrated an association between self-rated sensitivity to emotional and psychological issues in patients and the diagnosis of these issues in practice by chart audit.

However, both this study[25] and another[24] displayed variability among a subset of participants rating themselves as skilled. In the study by Robbins et al,[25] physicians who believed themselves to be better at detecting hidden emotions were in fact less accurate than their colleagues. In the study by Paradise et al,[24] there was wide variability among respondents and disagreement with expert consensus about the diagnosis of sexual abuse among up to 20% of self-styled experts.[24] Among those studies identified as demonstrating little or no relationship between self- and other-identified observations, there were 3 related findings regarding the misperceptions of abilities. First, Leopold et al[16] found an inverse relationship between confidence and competence in simulated joint injections. Second, Fox et al[15] demonstrated less concor-

Table 2. Studies Examining the Self-assessment of Practicing Physicians*

Source	Population; Participation Rate	Content Area; Domain of Self-assessment; Type of Self-assessment†	Self- and External Assessment Methods	Deficiencies in Methodological Quality; Study Outcomes‡
Amery and Lapwood,[13] 2004	17 UK general practioners with extra training in children's palliative care; 65%	Pediatric palliative care; Identification of learning needs; Concurrent	Physician questionnaire using 10-point Likert rating scale to measure confidence and learning needs in 19 domains of palliative care; Trained researchers reviewed physician-generated diary on unmet learning needs	Nonparticipants not described, no pilot testing of assessment measures, no statistical tests applied to outcomes; Physicians self-identified needs in biomedical areas but unable to identify 8 of 11 most highly rated needs (related to ethics, coping strategies, communication, and team functioning)
Leopold et al,[16] 2005	93 US practicing physicians, nurse practitioners, and physician assistants; 70%	Joint injection skills; Ability to assess confidence in performing procedural skills; Summative	Questionnaire using 10-point Likert rating scale to measure self-confidence in knee injection; Observed performance on a simulated knee with precise criteria for performance	Nonparticipants not described, no pilot testing of self- or external assessment, questionable use of Likert rating scale (use of 1 scale rather than sum of several scales); Significant inverse correlation with objective performance ($r = -0.25$, $P = .02$)§
Paradise et al,[24] 1997	414 US pediatricians, pediatric gynecologists; 38%	Detection of sexual abuse; Mental representation as "expert" compared with consensus standards; Summative	Physician self-rating as skilled or experienced vs less skilled or less experienced; Agreement with descriptions and interpretations of child sexual abuse case vignettes and photographs developed by expert consensus	No standard definition of expert, no pilot testing of vignettes; Self-rated experts matched expert opinion in 4 of 7 descriptions and in all 7 interpretations; Less skilled respondents identified 2 of 7 descriptions, 5 of 7 interpretations‖
Tracey et al,[27] 1997	67 New Zealand general practitioners; 67%	General medical knowledge in thyroid disease, sexually transmitted disease, and diabetes; Identification of learning needs; Concurrent	Physician questionnaire using a 9-point scale regarding assessment of knowledge in 20 general medical topics; True-false tests in 3 clinical areas, 2 of which were selected because of range of responses to needs assessment	No piloting testing of questionnaire; Correlations between self-assessment and performance on true-false test ranged between 0.19 and 0.21; $P = .11$ to $P = .15$¶
Woods et al,[28] 2004	178 US primary care and emergency medicine physicians#; 67%	Differential diagnosis of smallpox; Performance in case vignettes; Predictive	Physician rating on Likert scale of "comfort" with the diagnosis of smallpox; Performance based on differential diagnosis of vignettes and photographs	No pilot testing of questionnaire or vignettes, continuous outcomes dichotomized, and variable selection methods for logistic regression could have resulted in overfitting; Higher comfort scores correlated with higher differential diagnosis scores (OR, 2.2; 95% CI, 1.4-3.3)

Abbreviations: CI, confidence interval; OR, odds ratio.
*See Table 1 for studies by Claridge et al,[18] Fernandez et al,[19] and Young et al,[20] which also examined the self-assessment of practicing physicians.
†The 3 discrete types of self-assessment used in this review were predictive, summative, and concurrent, as defined in asterisk footnote in Table 1.
‡Unless otherwise indicated, all qualitative and descriptive measures were sufficiently described to permit replication. Most studies did not blind the external assessment to the self-rating (the study by Young et al[20] in Table 1 did blind the external assessment).
§Attending a course improved correlation between self- and external assessment, although overconfidence persisted.
‖There was wide variation among self-identified experts.
¶Some participants who scored well on the test indicated a lack of knowledge; others performed poorly on the test but rated themselves as having little or no learning needs in that area.
#Includes residents.

Table 3. Studies Examining the Self-assessment of Graduate Physician-Trainees and Practicing Physicians*

Source	Population; Participation Rate	Content Area; Domain of Self-assessment Type of Self-assessment†	Self- and External Assessment Methods	Deficiencies in Methodological Quality; Study Outcomes‡
Barnsley et al,[14] 2004	30 Australian junior medical officers (postgraduate year 1); 79%	Routine procedural skills; Future performance on competency (skills) tests; Predictive	Physician questionnaires measuring self-reported confidence for procedures on a 4-point Likert rating scale; Observed performance on 7-part observed structured clinical examination, pilot testing of criteria checklist, observers blinded	Null hypothesis not adequately stated for Wilcoxon test measures; Comparison of percentage scores between self-ratings and observed performance, observed structured clinical examination scores for all 7 dimensions were statistically significantly lower than self-reported: Wilcoxon *z* score ranged from –2.7 ($P<.01$) to –4.6 ($P<.001$)
Biernat et al,[12] 2003	12 US primary care residents; NA	Assessment of dementia; Ability to self-rate specific and global dimensions of care in recently completed standardized patient encounter; Summative	Physician self-completed 17-item checklist of performance in patient encounter; Global performance of patient encounter observed by experienced faculty and specific behavioral item analysis of patient visit by faculty supervisor	Sampling frame unclear, no statistical tests applied to specific item categories; Agreement reached on 6 of 7 items in global ratings, no agreement on any of the 17 specific items
Fox et al,[15] 2000	22 UK preregistration house officers (postgraduate year 1); 55%	Basic clinical skills; Performance compared with attending physicians' ratings and performance on an observed structured clinical examination; Predictive	Resident 5-point Likert rating scale to self-assess abilities in 15 common clinical skills; Rating by faculty supervisor of resident performance on pretested and validated observed structured clinical examination stations (scored by a trained observer)	Nonparticipants not described, no pilot testing of self-assessment method, no statistical tests applied to self- and supervisor ratings; Good agreement and little variation between residents' self-rating and supervisors, house officers tended to self-assess more accurately in stations in which they scored well and less able to do so in those stations in which they had difficulty; no significant correlations (Spearman *r*) between skills performed on observed structured clinical examination stations and participants' self-ratings
Hoppe et al,[21] 1990	54 US internal medicine and family practice residents; 100%	Counseling in primary disease prevention; Mental representation or perception of self as committed to health promotion practices; Summative	Resident questionnaire using 127 Likert-like rating scales to determine attitudes, opinions, and perceptions; Observed performance with 1 of 2 undetected standardized patients	No pilot testing of questionnaire or of standardized patients, no mention of detection rate of patients, pairing between performance measures and questionnaire items not defined; No significant ($P = .05$) correlations found between items on questionnaire and actual performance with standardized patients
Ireton and Sherman,[22] 1988	41 US final-year residents in family practice; 91%	Psychological aspects of medicine; Mental representations of self as interested and competent in managing the psychological aspects of primary care; Summative	Resident questionnaire using 11 Likert-like items; Rating of residents' abilities on a 3-point scale by 3 faculty members	Questionnaire previously used but with no established validity or reliability, 3-point scale for faculty ratings not described, statistical tests not well-defined; Significant correlations between faculty ratings and residents (*r* between 0.29 and 0.50; $P<.05$)
Johnson and Cujec,[23] 1998	24 Canadian residents in critical care (surgery, medicine, obstetrics, and anesthesia); 40%	Critical care medicine; Mental representation of self as competent clinician; Summative	Resident self-evaluation of overall competence and subsets of behavior§; Performance evaluation in 10 domains by nursing staff and supervising physicians and a multiple choice examination, multiple choice examination pilot tested	Participation rate in self-assessment formats by residents not described or explained; Self-evaluation vs nursing and physician staff evaluations correlated significantly in only 3 of 10 categories, no significant correlation between self-evaluations and multiple choice examinations
Parker et al,[17] 2004	311 US family medicine residents from 13 of 31 family medicine programs; NA	Basic medical knowledge; Scores on a future competency test; Predictive	Resident self-rating of anticipated performance on an in-training examination using a visual analog scale (converted to a 100-point scale); Actual score on an in-training examination	Total resident population not enumerated, self-assessment method not pilot tested; Pearson correlation coefficients were <0.3 in all categories of comparison‖
Robbins et al,[25] 1994	55 Canadian resident and staff primary care physicians; NA	Recognition of depression and anxiety in patient encounters; Mental representation as sensitive to patient emotional states; Summative	Resident questionnaire regarding sensitivity to hidden emotional states using standardized Likert rating scale; Patient depression scores on standardized tests and a chart review for psychological diagnoses by a blinded reviewer	Total number of residents and staff not given; Self-rated sensitivity to emotions and recognition of psychological distress in patient, physicians self-rating higher on this scale were significantly less accurate in their diagnoses

Abbreviation: NA, data not available.
*See Table 1 for study by Stern et al,[26] which examined the self-assessment of residents.
†The 3 discrete types of self-assessment used in this review were predictive, summative, and concurrent, as defined in asterisk footnote in Table 1.
‡Unless otherwise indicated, all qualitative and descriptive measures were sufficiently described to permit replication. Most studies did not blind the external assessment to the self-rating (the study by Young et al[20] in Table 1 did blind the external assessment).
§Evaluation forms developed by the American Board of Internal Medicine.
‖Residents in lowest quartile of examination score predicted their performance poorly.

dance in an OSCE setting between self- and other perceptions in poorly performed basic clinical skills than in those instances in which participants performed well. Third, Parker et al[17] found that residents scoring in the lowest quartile of a knowledge-based family practice examination recognized their learning needs less well than those in higher-achieving quartiles.

Regarding variables that might explain differences in accuracy of self-assessment, 2 studies reported the age[16,25] and 1 study reported the experience[28] of participants, linking this to self-assessment abilities. Age or experience did not correlate with the clinician's ability to judge performance in joint injections.[16] In contrast, age did correlate with a decreased propensity to diagnose emotional aspects of illness in one study[25] and with more accuracy in diagnosing smallpox in another.[28]

Methodological Quality

The studies demonstrated variability in methodological quality. The majority of populations studied were well described. Only 1 study[20] used the vague phrase "general practitioners." Additionally, the sampling frame was addressed in most studies, although the total number from which the sample was drawn was not identified in 3 studies[12,17,20] and a useful description of nonparticipants was not supplied in 9 studies.[13,15-18,23,25,26,28] The content domain was well characterized in all studies either in text or a tabular form. Two studies[13,20] used standard qualitative research methods such as tape-recorded and transcribed interviews. Only 1 study[20] referred to the blinding of an external observer.

Quantitatively, 9 studies[12,14-16,19,21,23,25,27] used pretested or validated measures such as OSCEs, standardized patients, and standardized instruments, and 2 of these studies[14,16] described objective criteria for performance observation. Two studies[14,15] used medical students for the pilot testing of the methods for comparison with self-assessment measures. In contrast, other studies[13,15-21,24,26,27] applied self- or other assessment instruments that were not described as having been pilot tested or validated. Regarding the use and reporting of statistical tests, we found examples of flawed methods. For example, 2 studies[12,22] did not report which tests were used. Other problems were insufficient or confusing data precluding confirmation of odds ratios,[19] dichotomizing variables,[19,28] confusion of and incomplete use of parameters in scales with inadequate statistical application,[20] and inadequate justification for use of statistics with mixed scales (eg, confidence comparisons with OSCE scores).[14,15] When multiple comparisons were performed, no description of efforts to control for inflation of type I error was provided.[12,18] No trends in improvement of methodological rigor were detected over the time span represented by these studies.

COMMENT

Relationship Between Self-rated Assessment and External Assessments

This systematic review found that in a majority of the relevant studies, physicians do not appear to accurately self-assess. Weak or no associations between physicians' self-rated assessments and external assessments were observed. While some studies found a reasonable association between physicians' demonstrated self-assessment abilities and external assessments in the area of cultural and linguistic sensitivity,[18] between self- and supervisor ratings at a general level,[12,15,22] between self- and external tests,[24,28] and between self-assessment and chart audit, wide variability and some errors in judgment are demonstrated in other studies.[24,25] In the studies indicating poor or limited accuracy of self-assessment, this finding was independent of level of training, specialty, the domain of self-assessment, or manner of comparison.

These findings are not new. Sibley et al[29] reported similar findings more than 2 decades ago, as did subsequent studies by Gordon[30] and Dunning et al.[31] The findings are consistent with studies in other disciplines. For example, in a meta-analysis of quantitative self-assessment studies in law, engineering, guidance counseling, behavioral science, psychology, and medicine, Falchikov and Boud[32] noted correlations between self- and external assessments of student performance ranging from 0.05 to 0.82, with a mean of 0.39. Within the health profession, Gordon[30] found that correlations for self-assessments of knowledge ranged from 0.02 to 0.65. Furthermore, despite our finding in 2 studies[24,28] that specific self-assessment may be reliable predictors of performance, Eva et al[33] found that poor correlations persist even when domains are well-defined. Finally, perhaps of greatest concern are the findings that those who perform the least well by external assessment also self-assess less well. These results have been demonstrated by others[34,35] and require further understanding.

Taken together, these conclusions prompt reflection on the use of self-rated assessment and its role in lifelong learning and value in regulation and patient care.

Construct and Study of Self-rated Assessment

These studies highlight several considerations for the study of self-assessment as an important domain of physician competence. First, the construct of self-rated assessment itself is not easily studied, in large part because its nature is neither fully developed nor tested.[10] We defined 3 discrete types of self-assessment: predictive, summative, and concurrent. The most value may come from attention to conceptual clarity and coherence for the field of self-assessment[36,37]; a more thorough understanding of continuous self-assessment; and a more precise definition of self-assessment to include an increased understanding of physicians' abilities to reflect[38,39] and of the nature of insight.[40] Furthermore, given that these studies shed only limited light on the process of self-assessment, fur-

ther research in this area might move beyond the boundaries of social and behavioral psychology to include cognitive, simulation, or other promising approaches such as the appraisal of perceived self-efficacy or examining the role that age and experience might play in the process.[41]

Second, if such studies of self-assessment are undertaken, researchers should augment study rigor and reportability by better describing their populations, sampling frame, and methods; more clearly differentiating between types of self-assessments; attempting to resolve questions of volunteer bias; and articulating best-practice templates for studying and reporting self-assessment compared with external assessment. In their review of self-assessment methods, Ward et al[36] also call for increased methodological rigor by improving the validity and reliability of the external reference standard, increasing the description of anchors in questionnaires used in self-assessment tools, and focusing more on reporting results for the individual and less on the group.

Limitations

Several limitations in this review should be considered. First, while literature searches were conducted by one of the authors and an independent information specialist to provide an exhaustive coverage of the literature, the lack of extensive Medical Subject Headings in the literature databases could have contributed to not retrieving some studies. We attempted to overcome this limitation by using proprietary search engines that used full-text search strategies. Second, some studies lacked full descriptions of methods, outcomes, and use of statistical tests, limiting our ability to describe the studies more fully, to develop explanatory hypotheses, or to generalize. Third, the domains of competence or performance in these studies, the tools used to measure them, and the assumptions (such as predictions of future performance on a test, or self-ratings in past performances as a teacher) were variable, precluding a meta-analytic approach.

Finally, it can be argued that the relatively small number of studies found in this review—with their mixture of methods, differing levels of physician training and experience, volunteer physician participation, and conceptual variability—provides an inadequate evidence base for understanding the ability of physicians to perform self-rated assessments. However, we believe that the selected literature offers fairly consistent evidence of the limited ability of physicians to independently assess their performance. These findings can help inform both further research in this area and the structure and practice of self-directed learning and self-assessment in graduate and continuing education.

Assessment Formats and Content

If it is true that physicians perform poorly in this domain, new initiatives and formats are needed to assist the self-assessment process and to more accurately promote and assess broader domains of competence such as professionalism and lifelong learning.[38] The positive findings were in global performance, with the potential for feedback on broad dimensions of care by faculty supervisors, given over time,[12,15,19,22] or in highly specialized areas such as child sexual abuse or bioterrorism,[24,28] in which the practitioner might be expected to accurately self-assess. Ultimately, a more useful approach may be to focus on externally determined self-assessments to guide the clinician in the use of educational and other activities designed to improve performance.

First, such measures might include the development of a more holistic continuing professional development process involving learning portfolios, documenting practice-based learning and improvement activities, creating less general and more detailed learning and practice objectives, and addressing the general competencies espoused by the Accreditation Council for Graduate Medical Education.[8]

Second, training may reduce the variation between self- and external assessments by encouraging the internalization of objective measurements or benchmarks of performance.[10] Although one study[16] in this review demonstrated only marginally improved correlation between confidence and skill performance following training and feedback, another study[42] reported that training increases the relationship between observed and self-reported hand-washing techniques following a hospital-wide quality improvement initiative. Similarly, physician trainees may be able to self-rate more accurately when they compare their ratings with those of others.[30] Attention to this effect of training and the comparative feedback phenomenon in undergraduate medical education as well as graduate and CME appears both appropriate and timely.[43,44]

Third, given that some improvement needs (eg, those in the psychosocial realm)[13,45,46] may be more difficult to self-assess, methods such as multisource feedback (360°) evaluations may be a necessary next step, particularly when interpersonal skills, communication skills, or professionalism needs to be evaluated.[47] Fourth, objective measures of competence and performance deserve serious consideration, especially when issues of medical licensure, recertification, quality, and patient safety are paramount. In this regard, the National Health Service in the United Kingdom has provided an example of externally informed self-assessment in formulating the concept of appraisal (the structured process of "facilitated self-reflection"[48]) in which an external appraiser guides and directs the process of self-assessment. Finally, specialty societies and others can increase their role in providing current evidence-based learning objectives on a regular basis to members of

their discipline, giving external markers of competence.

Author Contributions: Dr Davis had full access to all of the published data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Davis, Harrison, Mazmanian, Fordis.
Acquisition of data: Davis, Perrier.
Analysis and interpretation of data: Davis, Mazmanian, Fordis, Harrison, Thorpe.
Drafting of the manuscript: Davis, Fordis, Harrison, Mazmanian.
Critical revision of the manuscript for important intellectual content: Davis, Mazmanian, Fordis, Harrison, Thorpe, Perrier.
Statistical analysis: Mazmanian, Fordis, Thorpe.
Obtained funding: Davis, Perrier.
Administrative, technical, or material support: Perrier.
Study supervision: Davis.
Financial Disclosures: None reported.
Funding/Support: Dr Davis was supported in part by the Association of American Medical Colleges' Petersdorf scholar-in-residence program. Ms Perrier was supported in part by the Research and Development Resource Base in Continuing Medical Education, funded by the Academic Development Fund in Continuing Education of the University of Toronto, which is supported in part by the Alliance for Continuing Medical Education, the Society for Academic Continuing Medical Education, and the Royal College of Physicians and Surgeons of Canada.
Role of the Sponsor: No sponsors were involved in the design and conduct of the study; the collection, management and interpretation of the data; or the preparation, review and approval of the manuscript.
Acknowledgment: We acknowledge the support given by the Association of American Medical Colleges' Resource Center, especially Marian Talifero, in providing complimentary literature searches, and other members of a working group on Continuing Medical Education, including Nancy Davis, PhD. We also are grateful to Kevin Eva, PhD, McMaster University (Hamilton, Ontario), and Anton Kuzel, MD, MHPE, Virginia Commonwealth University (Richmond), for their comments. None of these individuals received any compensation for their assistance.

REFERENCES

1. Westberg J, Jason H. Fostering learners' reflection and self-assessment. *Fam Med*. 1994;26:278-282.
2. Davis NL, Willis CE. A new metric for continuing medical education credit. *J Contin Educ Health Prof*. 2004;24:139-144.
3. Johnson DA, Austin DL, Thompson JN. Role of state medical boards in continuing medical education. *J Contin Educ Health Prof*. 2005;25:183-189.
4. American Medical Association. Physician resources for CME. http://www.ama-assn.org/ama/pub/category/2922.html. Accessed March 23, 2006.
5. American Board of Medical Specialties. Approved initiatives for Maintenance of Certification for the ABMS board members. http://www.abms.org/Downloads/Publications/3-Approved%20Initiatives%20for%20MOC.pdf. Accessed June 9, 2006.
6. Wasserman SI, Kimball HR, Duffy FD; Task Force on Recertification. Recertification in internal medicine: a program of continuous professional development. *Ann Intern Med*. 2000;133:202-208.
7. Can MEDS 2000: extract from the CanMEDS 2000 Project Societal Needs Working Group Report. *Med Teach*. 2000;22:549-554.
8. Accreditation Council for Graduate Medical Education. General competencies. http://www.acgme.org/outcome/comp/compFull.asp. Accessed June 9, 2006.
9. Bashook PG, Miller SH, Parboosingh J, Horowitz SD, eds. Credentialing physician specialist: a world perspective proceedings. http://www.abms.org/Downloads/Conferences/Credentialing%20Physician%20Specialists.pdf. Accessed June 9, 2006.
10. Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med*. 2005;80:S46-S54.
11. Research and Development Resource Base in CME. http://www.cme.utoronto.ca/search. Accessed March 23, 2006.
12. Biernat K, Simpson D, Duthie E Jr, Bragg D, London R. Primary care residents self assessment skills in dementia. *Adv Health Sci Educ Theory Pract*. 2003;8:105-110.
13. Amery J, Lapwood S. A study into the educational needs of children's hospice doctors: a descriptive quantitative and qualitative survey. *Palliat Med*. 2004;18:727-733.
14. Barnsley L, Lyon PM, Ralston SJ, et al. Clinical skills in junior medical officers: a comparison of self-reported confidence and observed competence. *Med Educ*. 2004;38:358-367.
15. Fox RA, Ingham Clark CL, Scotland AD, Dacre JE. A study of pre-registration house officers' clinical skills. *Med Educ*. 2000;34:1007-1012.
16. Leopold SS, Morgan HD, Kadel NJ, Gardner GC, Schaad DC, Wolf FM. Impact of educational intervention on confidence and competence in the performance of a simple surgical task. *J Bone Joint Surg Am*. 2005;87:1031-1037.
17. Parker RW, Alford C, Passmore C. Can family medicine residents predict their performance on the in-training examination? *Fam Med*. 2004;36:705-709.
18. Claridge JA, Calland JF, Chandrasekhara V, Young JS, Sanfey H, Schirmer BD. Comparing resident measurements to attending surgeon self-perceptions of surgical educators. *Am J Surg*. 2003;185:323-327.
19. Fernandez A, Schillinger D, Grumbach K, et al. Physician language ability and cultural competence: an exploratory study of communication with Spanish-speaking patients. *J Gen Intern Med*. 2004;19:167-174.
20. Young JM, Glasziou P, Ward JE. General practitioners' self ratings of skills in evidence based medicine: validation study. *BMJ*. 2002;324:950-951.
21. Hoppe RB, Farquhar LJ, Stoffelmayr HR. Residents' attitudes towards and skills in counseling: using undetected standardized patients. *J Gen Intern Med*. 1990;5:415-420.
22. Ireton HR, Sherman M. Self-ratings of graduating family practice residents' psychological medicine abilities. *Fam Pract Res J*. 1988;7:236-244.
23. Johnson D, Cujec B. Comparison of self, nurse, and physician assessment of residents rotating through an intensive care unit. *Crit Care Med*. 1998;26:1811-1816.
24. Paradise JE, Finkel MA, Beiser AS, et al. Assessments of girls' genital findings and the likelihood of sexual abuse: agreement among physicians self-rated as skilled. *Arch Pediatr Adolesc Med*. 1997;151:883-891.
25. Robbins JM, Kirmayer LJ, Cathebras P, Yaffe MJ, Dworkind M. Physician characteristics and the recognition of depression and anxiety in primary care. *Med Care*. 1994;32:795-812.
26. Stern DT, Linzer M, O'Sullivan PS, Weld L. Evaluating medical residents' literature-appraisal skills. *Acad Med*. 1995;70:152-154.
27. Tracey JM, Arroll B, Richmond DE, Barham PM. The validity of general practitioners' self assessment of knowledge: cross sectional study. *BMJ*. 1997;315:1426-1428.
28. Woods R, McCarthy T, Barry MA, Mahon B. Diagnosing smallpox: would you know it if you saw it? *Biosecur Bioterror*. 2004;2:157-163.
29. Sibley JC, Sackett DL, Neufeld V, Gerrard B, Rudnick KV, Fraser W. A randomized trial of continuing medical education. *N Engl J Med*. 1982;306:511-515.
30. Gordon MJ. A review of the validity and accuracy of self-assessments in health professions training. *Acad Med*. 1991;66:762-769.
31. Dunning D, Heath C, Suls J. Flawed self-assessment: implications for health, education, and the workplace. *Psychol Sci Public Interest*. 2004;5:69-106.
32. Falchikov N, Boud D. Student self-assessment in higher education: a meta-analysis. *Rev Educ Res*. 1989;59:395-430.
33. Eva KW, Cunnington JP, Reiter HI, Keane DR, Norman GR. How can I know what I don't know? poor self assessment in a well-defined domain. *Adv Health Sci Educ Theory Pract*. 2004;9:211-224.
34. Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: novice physicians who are unskilled and unaware of it. *Acad Med*. 2001;76:S87-S89.
35. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol*. 1999;77:1121-1134.
36. Ward M, Gruppen L, Regehr G. Measuring self-assessment: current state of the art. *Adv Health Sci Educ Theory Pract*. 2002;7:63-80.
37. Colliver JA, Verhulst SJ, Barrows HS. Self-assessment in medical practice: a further concern about the conventional research paradigm. *Teach Learn Med*. 2005;17:200-201.
38. Epstein RM, Hundert EM. Defining and assessing professional competence. *JAMA*. 2002;287:226-235.
39. Mamede S, Schmidt HG. The structure of reflective practice in medicine. *Med Educ*. 2004;38:1302-1308.
40. Hays RB, Jolly BC, Caldon LJ, et al. Is insight important? measuring capacity to change performance. *Med Educ*. 2002;36:965-971.
41. Bandura A. Social cognitive theory: an agentic perspective. *Annu Rev Psychol*. 2001;52:1-26.
42. Moret L, Tequi B, Lombrail P. Should self-assessment methods be used to measure compliance with handwashing recommendations? a study carried out in a French university hospital. *Am J Infect Control*. 2004;32:384-390.
43. Simon FA, Aschenbrener CA. Undergraduate medical education accreditation as a driver of lifelong learning. *J Contin Educ Health Prof*. 2005;25:157-161.
44. Greiner AC, Knebel E, eds. *Health Professions Education: A Bridge to Quality*. Washington, DC: National Academy Press; 2003.
45. Sherman CD Jr, Davis DA. CME in oncology–from where we were to where we are going. *J Cancer Educ*. 1995;10:131-136.
46. Sachdeva AK. The new paradigm of continuing education in surgery. *Arch Surg*. 2005;140:264-269.
47. Lockyer J. Multisource feedback in the assessment of physician competencies. *J Contin Educ Health Prof*. 2003;23:4-12.
48. Conlon M. Appraisal: the catalyst of personal development. *BMJ*. 2003;327:3890-3891.

