Instruments for Evaluating Education in Evidence-Based Practice
A Systematic Review

Terrence Shaneyfelt, MD, MPH
Karyn D. Baum, MD, MSEd
Douglas Bell, MD, PhD
David Feldstein, MD
Thomas K. Houston, MD, MPH
Scott Kaatz, DO
Chad Whelan, MD
Michael Green, MD, MSc

Physicians often fail to implement clinical maneuvers that have established efficacy.\(^1,2\) In response, professional organizations have called for increased training in evidence-based practice (EBP) for all health care professions and at all levels of education.\(^3-6\) Evidence-based practice may be defined as the integration of the best research evidence with patients’ values and clinical circumstances in clinical decision making.\(^7\)

As educators implement EBP training, they need instruments to evaluate the programmatic impact of new curricula and to document the competence of individual trainees. Prior systematic reviews of EBP training summarized the effectiveness of educational interventions,\(^8-13\) but only 1 that was conducted in 1999 also included a detailed analysis of evaluation instruments.\(^8\) Although there are multiple components of EBP (BOX), as of 1998 the published instruments focused on critical appraisal to the exclusion of other EBP steps, mea-

See also Patient Page.
Box. Definitions of Variables and Terminology Used in This Study

Description: Format of instrument; choices include written or Web-based test, self-report survey, OSCE with standardized patients, other OSCE, portfolio, audiocassette of teaching sessions, record audit, chart-stimulated recall, direct observation (clinical evaluation exercise), rating scale, and other

Development: Free-text description of development

EBP domains

Knowledge: Knowledge about EBP

Skills: EBP skills are distinguished from knowledge by participants applying their knowledge by performing EBP steps in some type of clinical scenario, such as with a standardized patient, written case, computer simulation, OSCE, or direct observation.

Ask: Converting the need for information (about prevention, diagnosis, prognosis, therapy, causation, etc) into an answerable question

Acquire: Tracking down the best evidence with which to answer that question

Appraise: Critically appraising that evidence for its validity (closeness to the truth), impact (size of the effect), and applicability (usefulness in one's own clinical practice)

Apply: Applying the evidence in clinical decision making (includes both individualizing the evidence [such as recasting number needed to treat for the patient's baseline risk] and integrating the evidence with the patient's preferences and particular clinical circumstances)

Attitude: Attitudes toward EBP

Behaviors: Actual performance of EBP in practice

Enacting EBP steps in practice: Actually enacting EBP steps (such as identifying clinical questions) in the course of patient care activities

Performing evidence-based clinical maneuvers: Performing evidence-based maneuvers in trainee's actual practice, such as prescribing angiotensin-converting enzyme inhibitors for congestive heart failure with depressed left ventricular function or checking hemoglobin A1c in patients with diabetes

Affecting patient outcomes: Trainee's patients experience improved or favorable outcomes, such as lower blood pressure

Feasibility: Documentation of some measure of ease of implementation; choices include time required to administer instrument, time required to score instrument, expertise required to score instrument, cost to administer and score, administrative support required, other

Interrater reliability: Statistical test (κ or correlation coefficient) of the agreement among 2 or more raters' scoring of the responses. Applied only to instruments that required some level of judgment to score, such as free-text responses. In contrast, reliability testing was deemed not applicable for instruments that required no rater judgment to score, such as multiple-choice tests. Credited as "tested" if a quantitative assessment was done. Credited as "established" if the corresponding statistical test was significant.

Participants (number, discipline, and level): Participants in whom the instrument was tested; options include undergraduate medical students (year), residents (specialty), fellows (specialty), faculty physicians, practicing physicians, nurses in training, practicing nurses, allied health professionals, and other health care professionals

Validity: For all types except content validity, credited as "tested" if a quantitative assessment of a particular type of validity was done; credited as "established" if the corresponding statistical test was significant

Based on content: External review of the instrument by experts in EBP

Based on internal structure

Internal consistency: Statistical test to establish the relationship between items within either the entire instrument or a prespecified section of the instrument

Dimensionality: Factor analysis to determine if the instrument measured a unified latent construct or, if specified in advance, discrete subthemes

Based on relationship to other variables

Responsive: Ability to detect the impact of an EBP educational intervention; requires statistical comparison of same participant's scores before and after an EBP educational intervention

Discriminative: Ability to discriminate between participants with different levels of EBP expertise; requires statistical comparison of instrument scores among participants of different levels of EBP ability

Criterion: Statistical test of the relationship between the instrument scores and participants' scores on another instrument with established psychometric properties

Abbreviations: EBP, evidence-based practice; OSCE, observed structured clinical examination.

Classification of validity is based on the Standards for Educational and Psychological Testing of the Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education13 and other recommendations.15,16

sured EBP knowledge and skills but did not objectively document behaviors in actual practice, and often lacked established validity and reliability.8 In 2002, Hatala and Guyatt17 noted that “ironically, if one were to develop guidelines for how to teach [evidence-based medicine] based on these results, they would be based on the lowest level of evidence.” Since then, instruments have been developed to try to address the deficits in evaluation. In addition, EBP has become more sophisticated, requiring additional skills. For example, in identifying evidence, practi-
Figure. Search for and Selection of Articles for Review

| 2759 Potentially Relevant Articles Identified and Screened for Retrieval |
| 27192 Excluded (Not Reports of EBP Education or EBP Evaluation Instruments Based on Review of Title and/or Abstract) |
| 347 Selected for Full-Text Review |
| 252 MEDLINE |
| 225 EMBASE |
| 114 CINAHL |
| 13 ERIC |
| 8 HAPI |
| 115 Journal Table of Contents |
| 10 Internet Sites |

EBP indicates evidence-based practice. *Articles could be found in more than 1 database (see “Methods” section for details of search strategies, databases, and names of the 8 journals whose tables of contents were searched). †Reasons for exclusion not mutually exclusive.

tioners must be able to appraise, select among, and search emerging electronic secondary “preappraised” information resources. In applying evidence to decision making, they must explicitly integrate patient preferences and clinical context.

Because of these changes, we performed a systematic review of EBP evaluation instruments and strategies, documenting their development, format, learner levels, EBP evaluation domains, psychometric properties, and feasibility. Our 2 goals were to provide guidance for EBP educators by highlighting preferred instruments based on evaluation needs and to make recommendations for EBP education research based on the current state of the EBP evaluation science.

METHODS

Identification of Studies

To identify evaluation instruments, we searched the MEDLINE, EMBASE, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Health and Psychosocial Instruments (HAPI), and Educational Resources Information Center (ERIC) databases from 1980 through April 2006. Search terms included evidence-based medicine; critical appraisal; clinical epidemiology; journal club; clinical question; medical informatics; medical informatics applications; information storage and retrieval; databases, bibliographic; integrated advanced information management systems; MEDLARS; education; clinical trials; controlled clinical trials; multicenter studies; and program evaluation. We also manually searched the reference lists of retrieved articles, tables of contents of 8 major medical education journals (Academic Medicine, Medical Education, Teaching and Learning in Medicine, Medical Teacher, Advances in Health Sciences Education, Medical Education OnLine, Journal of Continuing Education in the Health Professions, and BioMed Central Medical Education), several EBP Internet sites, and the authors’ personal files. The Internet sites were chosen based on author experience as loci that might contain instruments not identified by other strategies.

We included studies that (1) reported an instrument or strategy that evaluated EBP knowledge, skills, attitudes, behaviors, or patient outcomes; (2) contained a sufficient description of the instrument or strategy to permit analysis; and (3) presented results of testing the performance of the instrument or strategy. We did not exclude any articles based on study design. Given the breadth of our review and the large number of articles initially captured by our search strategy, it was not feasible to translate the non–English-language articles to determine their suitability for inclusion. Thus, we limited our analysis to studies published in English. For 1 study, we contacted the authors for clarification. Studies that reported only satisfaction with a curriculum were excluded. Two authors (T.S. and M.G.) independently evaluated each article in the preliminary list for inclusion, and disagreements were resolved by consensus.

Data Extraction

We developed and piloted a standardized data form to abstract information from the included articles. A randomly assigned set of 2 raters, representing all permutations of the 6 raters, independently abstracted information from each of the included articles. In this process and in the article inclusion process, raters were not blinded to any portion of articles. After submitting their original abstraction forms to a central location, the pairs of raters resolved their differences by consensus. The abstraction variables included description and development of the EBP evaluation instrument; number, discipline, and training levels of participants; EBP domains evaluated; feasibility assessment; and type, method, and results of validity and reliability assessment (see Box for definitions).

We determined interrater reliability for the article inclusion process and for the data abstraction process based on data from all included articles. Statistics were calculated and interpreted according to the guidelines of Landis and Koch.

Quality Categorization of Studies

We did not use restrictive inclusion criteria related to study quality. However, we did define 3 levels of instruments, based on (1) the type, extent, methods, and results of psychometric testing and (2) suitability for different evaluation purposes. For use in the summative evaluation of individual trainees, we identified instruments with the most robust psychometric proper-
ties generally and, in particular, the ability to distinguish between participants of different levels of EBP experience or expertise (level 1). These instruments had to be supported by established interrater reliability (if applicable), objective (non–self-reported) outcome measures, and multiple (≥3) types of established validity evidence (including evidence of discriminative validity).

For use in evaluating the programmatic effectiveness of an EBP educational intervention, we identified a second group of instruments supported by established interrater reliability (if applicable) and “strong evidence” of responsive validity, established by studies with a randomized controlled trial or pre-post controlled trial design and an objective (non–self-reported) outcome measure (level 2). These instruments generally have less robust psychometric properties than level 1 instruments, which must be supported by 3 or more different types of validity evidence. However, level 2 instruments must be supported by higher-level (“strong”) evidence for responsive validity in particular. The criteria for “strong evidence” are stricter than the definition of responsive validity (Box) used for the general classifications in this review. Instruments meeting all of the criteria for level 1 may also have “strong evidence” for responsive validity (as indicated in the table footnotes) but this is not required for this designation.

Finally, considering the evaluation of EBP behaviors, we anticipated that few of the instruments would meet either of the preceding thresholds. Therefore, we used a single criterion of an objective (non–self-reported) outcome to distinguish a group of relatively high-quality measures in this domain (level 3).

In cases in which an instrument included 2 distinct pieces (with different formats) intended to evaluate 2 distinct EBP domains, we applied the quality criteria separately to each. For descriptive purposes, we included both subinstruments in the tables and indicated if one or both met the psychometric threshold.

We calculated descriptive statistics for the characteristics and psychometric properties of the evaluation instruments. Analyses were performed using Stata Special Edition version 9.0 (Stata Corp, College Station, Tex).

RESULTS

Inclusion criteria were met by 115 articles25-140 representing 104 unique assessment strategies (8 instruments were used in >1 study, and 1 study was reported in 2 articles) (FIGURE). There was substantial interrater agreement for the article inclusion process (κ = 0.68; 95% confidence interval [CI], 0.47-0.89), as well as for the assessments of validity based on content (κ = 0.70; 95% CI, 0.49-0.91) and based on internal structure (κ = 0.71; 95% CI, 0.47-0.95). There was moderate agreement on the assessment of validity based on relationships to other variables (κ = 0.52; 95% CI, 0.35-0.70).

Table 1. Characteristics of EBP Evaluation Instruments

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Instruments, No. (%) (N = 104)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participants’ health care profession discipline and training level</td>
<td></td>
</tr>
<tr>
<td>Students†</td>
<td>43 (41.5)</td>
</tr>
<tr>
<td>Postgraduate trainees‡</td>
<td>35 (33.7)</td>
</tr>
<tr>
<td>Practicing physicians</td>
<td>30 (28.8)</td>
</tr>
<tr>
<td>Nonphysicians‡</td>
<td>13 (12.5)</td>
</tr>
<tr>
<td>EBP evaluation domains</td>
<td></td>
</tr>
<tr>
<td>EBP knowledge</td>
<td>39 (37.5)</td>
</tr>
<tr>
<td>EBP skills</td>
<td>59 (56.7)</td>
</tr>
<tr>
<td>Ask</td>
<td>13 (12.5)</td>
</tr>
<tr>
<td>Acquire</td>
<td>35 (33.7)</td>
</tr>
<tr>
<td>Appraise</td>
<td>40 (38.5)</td>
</tr>
<tr>
<td>Apply</td>
<td>13 (12.5)</td>
</tr>
<tr>
<td>EBP attitudes</td>
<td>27 (26.0)</td>
</tr>
<tr>
<td>EBP behaviors</td>
<td>30 (28.8)</td>
</tr>
<tr>
<td>Performing EBP steps in practice</td>
<td>34 (32.7)</td>
</tr>
<tr>
<td>Performing evidence-based clinical maneuvers in practice</td>
<td>3 (2.9)</td>
</tr>
<tr>
<td>Patient outcomes</td>
<td>2 (1.9)</td>
</tr>
</tbody>
</table>

Abbreviation: EBP, evidence-based practice.
*See Box for definitions. Categories are not mutually exclusive.
†Medical students (n = 43), dental students (n = 1), and nursing students (n = 1).
‡Internal medicine (n = 19), emergency medicine (n = 1), surgery (n = 2), obstetrics/gynecology (n = 3), pediatrics (n = 1), and family medicine (n = 8) residents.
§Nurses (n = 7), physical therapists (n = 1), researchers (n = 1), and not specified (n = 4).

Characteristics of EBP Evaluation Instruments

The participants’ health care professions discipline and training level and the evaluated EBP domains are shown in Table 1 (see Box for definitions). The majority of instruments targeted students and postgraduate trainees, while nonphysicians were rarely evaluated. The instruments most commonly evaluated EBP skills (57%), followed by knowledge and behaviors (both 38%), followed by attitudes (26%). Among the EBP skills, critical appraisal of evidence was included in the greatest proportion of instruments.

Thirty (86%) of the 35 evaluation approaches for the “acquire” step related exclusively to skills in searching MEDLINE or similar bibliographic databases for original articles. Of the 5 instruments considering alternative electronic information sources, 4 specifically evaluated awareness, preference for, or skills in searching specific secondary evidence-based medical information resources (including the Cochrane Library, Database of Abstracts of Reviews of Effectiveness,ACP Journal Club, and Clinical Evidence)25,33,43,135 while the remaining one42 merely referred to “Web sites.” Similarly, among the instruments evaluating the “apply” step, only 5 (38%) of 13 went beyond the ability to consider research evidence to also assess the ability to integrate the evidence with the patient’s particular clinical context and preferences. Evaluation approaches included standardized patient ratings of students explaining a therapeutic decision after reviewing research evidence.38,39 Scoring of residents’ free-text justification of applying results of a study to a “paper case,”28 and documenting decision making before and after access to a research abstract11 or MEDLINE search.40

Most of the instruments evaluating EBP behaviors measured the use of EBP steps in practice. Of these, only 6 (18%) of 34 used objective outcome measures31,52,54,113,137 with the remaining relying on retrospective self-
established, followed by discriminative validity was most commonly tested and established. Among these, responsive validity evidence were established for only 34% of the 104 EBP evaluation instruments.26,39,60,100,114 Investigators conducted at least 1 type of interrater reliability testing on 21 (41.2%) of the 51 instruments for which it was appropriate, most commonly using k statistics and correlation coefficients.

Investigators conducted at least 1 type of validity testing in 64% and established it in 53% of the 104 EBP evaluation instruments (Table 2). However, multiple (≥3) types of validity evidence were established for only 10% of the instruments. Investigators most commonly sought (57%) and established (44%) evidence for validity based on relationships to other variables. Among these, responsive validity was most commonly tested and established, followed by discriminative and criterion validity.

Eight instruments were used in subsequent studies, either for further validation or to evaluate programmatic impact of an EBP curriculum. One instrument60 was used in 3 later studies61-63; 1 instrument140 was used in 2 later studies33,41; and 6 instruments33,41,59,64,133,136 were used in 1 subsequent study each.26,65,87,134,136,139

Feasibility and Psychometric Testing

Feasibility of implementation was reported for 19 (18.3%) of the 104 instruments. Among these, 13 reported the time required to administer or score the instrument,4 4 described the expertise required for scoring,37,47,65,72 and 4 estimated the financial costs of implementation.28,54,100,114 Investigators performed interrater reliability testing on 21 (41.2%) of the 51 instruments for which it was appropriate, most commonly using k statistics and correlation coefficients.

Eight instruments were used in subsequent studies, either for further validation or to evaluate programmatic impact of an EBP curriculum. One instrument60 was used in 3 later studies61-63; 1 instrument140 was used in 2 later studies33,41; and 6 instruments33,41,59,64,133,136 were used in 1 subsequent study each.26,65,87,134,136,139

Quality Categorization of Instruments

Level 1 Instruments. Table 3 summarizes the EBP evaluation domains, format, and psychometric properties of the instruments supported by established interrater reliability (if applicable), objective (non–self-reported) outcome measures, and multiple (≥3) types of established validity evidence (including evidence for discriminative validity). These instruments are distinguished by the ability to discriminate between different levels of expertise or performance and are therefore suited to document the competence of individual trainees. Furthermore, the robust psychometric properties in general support their use in formative or summative evaluations. The Fresno Test25 and Berlin Questionnaire59 represent the only instruments that evaluate all 4 EBP steps. In taking the Fresno Test, trainees perform realistic EBP tasks, demonstrating applied knowledge and skills. However, more time and expertise are required to grade this instrument. The multiple-choice format of the Berlin Questionnaire restricts assessment to EBP applied knowledge but also makes it more feasible to implement. The other instruments in Table 3 evaluate a narrower range of EBP.

Level 2 Instruments. In addition to 4 of the instruments in Table 3,26,39,60,64 9 instruments fulfilled the criteria for strong evidence of responsive validity (Table 4). These are appropriate to consider for evaluating programmatic (rather than individual) impact of EBP interventions. Six evaluated EBP knowledge and skills.27-31 Among these, only one27 measured all 4 EBP steps. Residents articulated clinical questions, conducted MEDLINE searches, performed calculations, and answered free-text questions about critical appraisal and application of the evidence. In this study, gains in skills persisted on re-testing at 6 months, indicating both concurrent and predictive responsive validity. The instrument described by Green and Ellis58 required free-text responses about the appraisal of a re-dacted journal article and application of the results to a patient. The 3 multiple-choice tests30-31 detected improvements in trainees’ EBP knowledge. However, in 2 of the studies, this gain did not translate into improvements in critical appraisal skills as measured with a test article59 or the incorporation of literature into admission notes.30 Finally, in Villanueva et al,37 librarians identified elements of the patient-intervention-comparison-outcome (PICO) format141 in clinical question requests, awarding 1 point for each included element. In a randomized controlled trial of instruction in clinical question construction, this instrument detected improvements in this skill.

Four EBP behavior instruments met the criteria for strong evidence of responsive validity and an objective outcome measure.31,52,73,113 Among these, 3 measured the enactment of EBP steps in practice.31,52,113 Ross and Verdieck31 analyzed audiotapes of resident-faculty interactions, looking for phrases related to literature searching, clinical

---

Table 2. Psychometric Characteristics of Evidence-Based Practice Evaluation Instruments

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Tested (N = 104)</th>
<th>Established</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content validity</td>
<td>17 (16.3)</td>
<td>17 (16.3)</td>
</tr>
<tr>
<td>Validity based on internal structure</td>
<td>13 (12.5)</td>
<td>13 (12.5)</td>
</tr>
<tr>
<td>Internal consistency</td>
<td>13 (12.5)</td>
<td>13 (12.5)</td>
</tr>
<tr>
<td>Dimensionality</td>
<td>6 (5.8)</td>
<td>6 (5.8)</td>
</tr>
<tr>
<td>Validity based on relationship to other variables</td>
<td>59 (56.7)</td>
<td>46 (44.2)</td>
</tr>
<tr>
<td>Responsive validity</td>
<td>51 (49.0)</td>
<td>41 (39.4)</td>
</tr>
<tr>
<td>Discriminative validity</td>
<td>10 (9.6)</td>
<td>9 (8.7)</td>
</tr>
<tr>
<td>Criterion validity</td>
<td>7 (6.7)</td>
<td>4 (3.9)</td>
</tr>
<tr>
<td>Instruments with ≥3 types of validity tests</td>
<td>11 (10.6)</td>
<td>10 (9.6)</td>
</tr>
</tbody>
</table>

*See Box for validity definitions. Categories are not mutually exclusive.
Table 3. Level 1 Instruments (Individual Trainee Formative or Summative EBP Evaluation)*

<table>
<thead>
<tr>
<th>Source</th>
<th>Study Settings/Participants</th>
<th>EBP Domains</th>
<th>Description</th>
<th>Inter-rater Reliability†</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ramos et al§ (Fresno Test)</td>
<td>53 “Experts” and 43 family practice residents and faculty in instrument development study</td>
<td>Formulate a focused question</td>
<td>Open-ended free-text questions, fill-in-the-blank questions, and calculations relating to 2 pediatric clinical scenarios; scored using a standardized grading rubric that includes examples of acceptable answers and specifies 4 or 5 grading categories (not evident, minimal and/or limited, strong, excellent) each of which is associated with a point value</td>
<td>Yes (R = 0.72-0.96)</td>
<td>Content Discriminative</td>
</tr>
<tr>
<td>Bennett et al§</td>
<td>79 Medical students in various clerkships in pre-post controlled trial</td>
<td>Critical appraisal skills</td>
<td>Set of case-based problems that require a diagnostic or treatment decision matched with an article advocating the test or treatment; students must “take a stand” and “defend” it in writing; graded on preset criteria.</td>
<td>Yes (κ = 0.74-1.00)†</td>
<td>Content Discriminative</td>
</tr>
<tr>
<td>Fritsche et al§, Aki et al§§ (Berlin Questionnaire)§</td>
<td>43 “Experts,” 20 third-year students, 203 participants in EBP course in instrument development study§§</td>
<td>Knowledge about interpreting evidence</td>
<td>2 Separate sets of 15 multiple-choice questions built around “typical” clinical scenarios</td>
<td>NA</td>
<td>Content Discriminative</td>
</tr>
<tr>
<td>Taylor et al§, Bradley and Henrin§, Bradley et al§§</td>
<td>152 Health care professionals in instrument development study§§</td>
<td>Knowledge of critical appraisal</td>
<td>Sets of 6 multiple-choice questions with 3 potential answers, each requiring a true, false, or “don’t know” response; best score on each set = 18</td>
<td>NA</td>
<td>Content Discriminative</td>
</tr>
<tr>
<td>MacRae et al§§, §§</td>
<td>44 Surgery residents in instrument development study§§</td>
<td>Critical appraisal skills</td>
<td>3 Journal articles, each followed by a series of short-answer questions and 7-point scales to rate the quality of elements of the study design; short-answer questions based on cards from an EBP textbook§§.</td>
<td>Yes (R = 0.78-0.91)</td>
<td>Internal consistency Discriminative</td>
</tr>
<tr>
<td>Weberschock et al§</td>
<td>132 Third-year medical students and 11 students with advanced training in “EBM working group” in development and pre-post uncontrolled study of peer-teaching EBP curriculum§</td>
<td>EBP knowledge and skills (specific EBP steps not specified)</td>
<td>5 Sets of 20 multiple-choice questions; (5 “easy,” 10 “average,” and 5 “difficult”) linked to clinical scenarios and pertaining to data from published research articles</td>
<td>NA</td>
<td>Internal consistency Discriminative</td>
</tr>
<tr>
<td>Haynes et al§§, §§ McKibbon et al§§</td>
<td>308 Physicians and physicians in training in RCT of one-one precepting and searching feedback§§</td>
<td>MEDLINE searching skills</td>
<td>Search output scored by comparison to searches for the same clinical questions by an expert end user physician and a librarian; “relative recall” calculated as number of relevant citations from a given search divided by number of relevant citations from the 3 searches (participant, expert physician, and librarian); “precision” calculated as the number of relevant citations retrieved in a search divided by the total citations retrieved in that search; article “relevance” rated reliably on a 7-point scale</td>
<td>Yes (κ = 0.79)†</td>
<td>Content Discriminative</td>
</tr>
</tbody>
</table>

Abbreviations: EBP, evidence-based practice; NA, not applicable; RCT, randomized controlled trial.
*These instruments had to be supported by established inter-rater reliability (if applicable), objective (non-self-reported) outcome measures, and 3 or more types of established validity evidence (including evidence for discriminative validity).
†Reliability testing was deemed not applicable for instruments that required no rater judgment to score, such as multiple-choice tests (see Box).
‡Demonstrated both interrater and intrarater reliability.
§Instruments evaluated in more than 1 study. Results from all of the studies were used to determine number of trainees, reliability, and validity.
¶Met level 1 criteria for searching skill portion of overall instrument, not for the searching behavior portion.

©2006 American Medical Association. All rights reserved. (Reprinted) JAMA, September 6, 2006—Vol 296, No. 9 1121

Downloaded from www.jama.com at ARCISPEDALE HOSPITAL, on October 5, 2006
epidemiology, or critical appraisal. Family practice residents’ “evidence-based medicine utterances” increased from 0.21 per hour to 2.9 per hour after an educational intervention. Stevermer et al.133 questioned residents about their awareness and knowledge of findings in recent journal articles relevant to primary care practice. Residents exposed to academic detailing recalled more articles and correctly answered more questions about them. Focusing on the “acquire” step, Cabell et al.132 electronically captured trainees’ searching behaviors, including number of logs to databases, searching volume, abstracts or articles viewed, and time spent searching. These measures were responsive to an intervention including a 1-hour didactic session, use of well-built clinical question cards, and prac-

### Table 4. Level 2 Instruments (Programmatic EBP Curriculum Evaluation)

<table>
<thead>
<tr>
<th>Source</th>
<th>Study</th>
<th>Settings/Participants</th>
<th>Knowledge and Skill Domains</th>
<th>Description</th>
<th>Interrater Reliability</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smith et al.37</td>
<td>55 Medical residents in pre-post controlled crossover trial of 7-week EBP curriculum, which included interactive sessions and computer laboratory training</td>
<td>Skills in formulating clinical questions Skills in MEDLINE searching Skills in critical appraisal Skills in applying evidence to individual patient decision making Knowledge of quantitative aspects of diagnosis and treatment studies</td>
<td>Test including sets of questions (format not specified) relating to 5 clinical cases</td>
<td>Yes (not reported)</td>
<td>Responsive‡</td>
<td></td>
</tr>
<tr>
<td>Green and Ellis39</td>
<td>34 Residents in controlled trial of a 7-session EBP curriculum</td>
<td>Skills in critical appraisal Skills in applying evidence to individual patient decision making</td>
<td>9-Question test (requiring free-text response) relating to a case presentation and a reprinted journal article</td>
<td>Yes (F² = 0.87)</td>
<td>Content Responsive</td>
<td></td>
</tr>
<tr>
<td>Linzer et al.40</td>
<td>44 Medical residents in RCT of journal club curriculum</td>
<td>Epidemiology and biostatistics knowledge Skills in critical appraisal</td>
<td>Multiple-choice test (knowledge); 15 questions chosen so that perfect score would allow access to 81% of medical literature25</td>
<td>NA</td>
<td>Content</td>
<td></td>
</tr>
<tr>
<td>Linzer et al.40</td>
<td>44 Medical residents in RCT of journal club curriculum</td>
<td>Epidemiology and biostatistics knowledge Skills in critical appraisal</td>
<td>Free-text critical appraisal of text article; scoring based on “gold standard” criteria developed by consensus of faculty</td>
<td>Yes (difference in observer variability = 1.1)</td>
<td>Content Discriminative</td>
<td></td>
</tr>
<tr>
<td>Landry et al.41</td>
<td>146 Medical students in controlled trial of 2 90-minute seminars</td>
<td>Research design and critical appraisal knowledge Skills in applying medical literature to critical appraisal service</td>
<td>10-Item test Blinded review of patient “write-ups” looking for literature citations</td>
<td>NA</td>
<td>Content Responsive None</td>
<td></td>
</tr>
<tr>
<td>Moss and Verdecia42</td>
<td>48 Family practice residents in controlled trial of 10-session EBP workshop (control residents in different program)</td>
<td>EBP knowledge (specific steps not specified) EBP behavior (enacting EBP steps in practice)</td>
<td>60-Item “open-book” multiple-choice test Analysis of audiotapes of resident-faculty interactions looking for phrases related to literature searching, clinical epidemiology, or critical appraisal</td>
<td>NA</td>
<td>Content Responsive Content Responsive</td>
<td></td>
</tr>
<tr>
<td>Villanueva et al.43</td>
<td>39 Health care professional participants in a library “evidence search and critical appraisal service” in an RCT of providing instructions and clinical question examples</td>
<td>Skills in formulating clinical questions</td>
<td>Librarians identified elements of the patient-intervention-comparison-outcome (PICO) format in clinical question requests141; 1 point awarded for each of 4 elements included</td>
<td>Yes (κ = 0.68)</td>
<td>Responsive</td>
<td></td>
</tr>
<tr>
<td>Cabell et al.44</td>
<td>48 Internal medicine residents in RCT of EBP curriculum, which included a 1-hour didactic session, the use of well-built clinical question cards, and practical sessions in clinical question building</td>
<td>EBP behavior (performance of EBP steps in practice)</td>
<td>Library system electronically captures MEDLINE searching behavior, including number of log-ons, searching volume, abstracts viewed, full-text articles viewed, and time spent searching</td>
<td>NA</td>
<td>Content</td>
<td>Responsive</td>
</tr>
<tr>
<td>Langham et al.45</td>
<td>Primary care practice teams in RCT of practice-based training in EBP and/or patient information management</td>
<td>EBP behaviors (performing evidence-based clinical maneuvers) EBP behaviors (affecting patient outcomes)</td>
<td>Record audit for physician performance and patient-level quality indicators relating to cardiovascular risk reduction</td>
<td>NA</td>
<td>Content</td>
<td></td>
</tr>
<tr>
<td>Stevermer et al.46</td>
<td>59 Family practice residents in RCT of EBP academic detailing</td>
<td>EBP behavior (performance of EBP steps in practice)</td>
<td>Test of awareness and recall of recently published articles reporting “important findings about common primary care problems” (selected by faculty physicians)</td>
<td>NA</td>
<td>Content</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: EBP, evidence-based practice; NA, not applicable; RCT, randomized controlled trial.

*Instruments supported by established interrater reliability (if applicable), “strong evidence” of responsive validity, established by studies with a randomized controlled trial or pre-post controlled trial design, and an objective [non–self-reported] outcome measure. Four instruments from Table 3 also had “strong evidence” of responsive validity and are appropriate for similar uses.25,30,36,62,65

†Interrater testing was deemed not applicable for instruments that required no rater judgment to score, such as multiple-choice tests (see Box).

‡Gains in skills persisted after 6 months, indicating both concurrent and predictive (responsive) validity.

§Met level 2 criteria for the EBP knowledge portion of overall instrument, not for the EBP skill portion.

| Quality indicators included recording of serum cholesterol and changes in serum cholesterol levels, recording of blood pressure and blood pressure control, recording of smoking status, and aspirin use. |
tactical sessions in clinical question building. One EBP behavior instrument in this category evaluated EBP practice performance and patient outcomes using medical record audits. Langham et al\(^5\) evaluated the impact of an EBP curriculum, documenting improvements in practicing physicians’ documentation, clinical interventions, and patient outcomes related to cardiovascular risk factors.

Although 3 controlled studies demonstrated the responsive validity of having librarians score MEDLINE search strategies\(^34,36\) or clinical question formulations\(^38\) according to predetermined criteria, these did not meet the criteria for interrater reliability testing.

**Level 3 Instruments.** In addition to the 5 EBP behavior instruments included in levels 1 and 2,\(^31,52,57,113,137\) 4 others used objective outcome measures but did not demonstrate strong evidence of responsive validity or multiple sources of validity evidence (Table 5).\(^55,54,58,140\) Two of these consisted of electronic learning portfolios that allowed trainees to document their enactment of EBP steps.\(^53,54\)

The remaining 2 instruments measured the performance of evidence-based maneuvers or patient outcomes. Ellis et al\(^56\) devised a reliable method for determining the primary therapeutic intervention chosen by a practitioner and classifying the quality of evidence supporting it. In this scheme, interventions are (1) supported by individual or systematic reviews of randomized controlled trials, (2) supported by “convincing nonexperimental evidence,” or (3) lacking substantial evidence. This instrument was subsequently used in 2 pre-post (but uncontrolled) studies of EBP educational interventions.\(^55,56\) Finally, Epling et al\(^58\) performed a record audit before and after residents developed and implemented a diabetes clinical guideline.

**Comment**

We found that instruments used to evaluate EBP were most commonly administered to medical students and postgraduate trainees and evaluated skills in searching for and appraising the evidence. At least 1 type of validity evidence was demonstrated in 53% of instruments (most commonly based on relationship to other variables), but multiple types of validity evidence were established for very few.

Educators need instruments to document the competence of individual trainees and to evaluate the programmatic impact of new curricula. Given the deficits of instruments previously

---

**Table 5. Level 3 Instruments (EBP Behavior Evaluation)**\(^*\)

<table>
<thead>
<tr>
<th>Source</th>
<th>Study Settings/Participants</th>
<th>Knowledge and Skill Domains</th>
<th>Description</th>
<th>Interrater Reliability†</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crowley et al(^57)</td>
<td>82 Medical residents in prospective cohort study</td>
<td>EBP behavior (enacting EBP steps in practice)</td>
<td>Internet-based portfolio (&quot;compendium&quot;) that allows residents to enter their clinical questions, information searching (via MEDLINE reference links), appraisal of articles, and impact on patient care decisions</td>
<td>NA</td>
<td>None</td>
</tr>
<tr>
<td>Fung et al(^54)</td>
<td>41 Obstetrics/gynecology residents across 4 programs in prospective cohort study</td>
<td>EBP behavior (enacting EBP steps in practice)</td>
<td>Internet-based learning portfolio that allows residents to describe an initiating clinical scenario, enter their clinical questions, link to information resources, and document learning points and implications for their practice</td>
<td>NA</td>
<td>None</td>
</tr>
<tr>
<td>Straus et al(^55), Lucas et al(^56), Ellis et al(^56)</td>
<td>33 Inpatient internal medicine physicians and 12 residents in pre-post uncontrolled trial of multicomponent EBP curriculum, which included 7 one-hour sessions and provision of evidence-based resources on the hospital network(^56)</td>
<td>EBP behaviors (performing evidence-based clinical maneuvers)</td>
<td>Record audit to determine the “primary therapeutic intervention” chosen by a practitioner and rate the level of evidence supporting it: (1) supported by individual or systematic reviews of RCTs, (2) supported by “convincing nonexperimental evidence,” or (3) lacking substantial evidence</td>
<td>Yes</td>
<td>Responsive (R = 0.76-0.92); (R = 0.20-0.56); not reported(^56)</td>
</tr>
<tr>
<td>Epling et al(^58)</td>
<td>11 Family practice residents in pre-post trial of EBP curriculum that included development of a clinical guideline</td>
<td>EBP behaviors (performing evidence-based clinical maneuvers)</td>
<td>Record audit for physician performance and patient-level quality indicators relating to care of patients with diabetes mellitus(^58)</td>
<td>NA</td>
<td>None</td>
</tr>
</tbody>
</table>

Abbreviations: EBP, evidence-based practice; NA, not applicable; RCT, randomized controlled trial.

\(^*\)EBP behavior instruments with objective (non–self reported) outcome measures. Five EBP behavior instruments from previous tables also have objective outcome measures and are appropriate for similar uses.\(^31,52,57,113,137\)

\(^†\)Reliability testing was deemed not applicable for instruments that required no rater judgment to score, such as multiple-choice tests (see Box).

\(^\dagger\)Instrument evaluated in more than 1 study. Results from all of the studies were used to determine number of trainees, reliability, and validity.

\(\S\)Quality-of-care indicators for diabetes included blood pressure measurement, fingerstick glucose measurement, hemoglobin A\(_1\)c control, documentation of foot examinations, and referral to nutritionists.

©2006 American Medical Association. All rights reserved.
available, it is not surprising that in 2000 only a minority of North American internal medicine programs objectively evaluated the effectiveness of their EBP curricula.\textsuperscript{143} Currently, there is a much wider selection of instruments, some of which are supported by more robust psychometric testing. While, like their predecessors, the currently available instruments most commonly evaluate critical appraisal, many more also measure the other important EBP steps. Among the instruments evaluating EBP behaviors, most continue to measure the performance of EBP steps by self-report. However, new instruments objectively document EBP steps and document the performance of evidence-based clinical maneuvers.

The choice of an EBP evaluation instrument should be guided by the purpose of the evaluation and the EBP domains of interest. The instruments in Table 3 are appropriate for evaluating the competence of individual trainees. Although they have reasonably strong psychometric properties, we believe that in the absence of well-defined passing standards for different learner levels, they should not yet be used for high-stakes evaluations, such as academic promotion or certification.

To evaluate the programmatic impact of EBP educational interventions, educators may turn to instruments with strong evidence of responsive validity (Table 4) and whose evaluation domains correspond with the objectives of their curricula. A conceptual framework for evaluating this aspect of EBP teaching has been developed by the Society of General Internal Medicine Evidence-Based Medicine Task Force.\textsuperscript{144} It recommends considering the learners (including their level and particular needs), the intervention (including the curriculum objectives, intensity, delivery method, and targeted EBP steps), and the outcomes (including knowledge, skills, attitudes, behaviors, and patient-level outcomes). With the exception of the instruments also included in Table 3, educators should use caution in using these instruments to assess the EBP competence of individual trainees because they were developed to evaluate the effectiveness of specific curricula and lack evidence for discriminative validity.

Only 5 EBP behavior instruments met the 2 highest quality thresholds in our analysis. Notwithstanding the psychometric limitations, however, it is important to document that trainees apply their EBP skills in actual practice. Our review identified several studies that documented EBP behaviors through retrospective self-report. However, this approach may be extremely biased, as physicians tend to underestimate their information needs and overestimate the degree of their pursuit.\textsuperscript{145} We recommend that educators restrict their selection of instruments to those with objectively measured outcomes.

Regarding the enactment of EBP steps in practice, analyzing audiotapes of teaching interactions\textsuperscript{31} and electronically capturing searching behavior\textsuperscript{32} showed responsive validity. However, we believe that these approaches fail to capture the pursuit and application of information in response to particular clinical questions, rendering them poor surrogates for EBP behaviors. Evidence-based practice learning portfolios,\textsuperscript{33,34} which serve as both an evaluation strategy and an educational intervention, may represent the most promising approach to document the performance of EBP steps. However, their use in any assessment with more serious consequences than formative evaluation must await more rigorous psychometric testing.

In addition to documenting the performance of EBP steps, educators are charged with documenting behavioral outcomes of educational interventions further downstream, such as performance of evidence-based clinical maneuvers and patient-level outcomes.\textsuperscript{146} The reliable approach of rating the level of evidence supporting clinical interventions has been widely used.\textsuperscript{140} In 2 studies, this approach detected changes following an EBP curriculum\textsuperscript{35} or supplying physicians with a literature search\textsuperscript{36} but, in the absence of controlled studies, did not meet our threshold for strong evidence of responsive validity. This system appears most suited to evaluating changes in EBP performance after an educational intervention or over time. To use it to document an absolute threshold of performance would require knowing the “denominator” of evidence-based therapeutic options for each trainee’s set of patients, making it impractical on a programmatic scale. The performance of evidence-based maneuvers may also be documented by auditing records for adherence to evidence-based guidelines or quality indicators. Hardly a new development, this type of audit is commonly performed as part of internal quality initiatives or external reviews. Our review found 2 examples of quality audits used to evaluate the impact of EBP training.\textsuperscript{37,38}

Assessing EBP attitudes may uncover hidden but potentially remediable barriers to trainees’ EBP skill development and performance. However, while several instruments contain a few attitude items, few instruments assess this domain in depth.\textsuperscript{33,36,31,133} Moreover, no attitude instruments in this review met our quality criteria for establishment of validity. One instrument demonstrated responsive validity in an uncontrolled study\textsuperscript{13} and another demonstrated criterion validity in comparison with another scale.\textsuperscript{50}

There are limitations that should be considered in interpreting the results of this review. As in any systematic review, it is possible that we failed to identify some evaluation instruments. However, we searched multiple databases, including those containing unpublished studies, using a highly inclusive search algorithm. Because our search was limited to English-language journals, we would not capture EBP instruments described in other languages. This might introduce publication bias if such instruments differ systematically from those appearing in English-language journals. Our exclusion of insufficiently described instruments may have biased our analysis if these differed systematically from the others. Our abstraction process showed
EVALUATING EDUCATION IN EVIDENCE-BASED PRACTICE

Disclaimer: Drs Baum and Green have published articles that are cited as part of this review. Neither abstracted data from their own published works.

Acknowledgment: This study was conducted as part of a charge of the Society of General Internal Medicine Evidence-Based Medicine Task Force, of which 3 authors (Drs Shaneyfelt, Whelan, and Green) are members. We thank Heather Coley, MPH, Department of Medicine, University of Alabama School of Medicine, for her assistance in compiling data on the citation of articles in various electronic databases.

REFERENCES

44. Bergus GR, Emerson M. Family medicine residents do not ask better-formulated clinical questions as they advance in their training. Fam Med. 2005;37:486-490.
EVALUATING EDUCATION IN EVIDENCE-BASED PRACTICE


112. Steele G, Greenidge E. Integrating medical communication skills with library skills curricula to teach health professionals to search MEDLINE. *J Med Libr Assoc.* 1987;5:50-51.


