Validation of the Fresno test of competence in evidence based medicine

Kathleen D Ramos, Sean Schafer, Susan M Tracz

Abstract

Objective To describe the development and validation of a test of knowledge and skills in evidence based medicine.

Design Cross sectional study.

Setting Family practice residency programme in California; a list server for those who teach evidence based medicine; and an evidence based medicine seminar series.

Participants Family practice residents and faculty members (n=43); volunteers self identified as experts in evidence based medicine (n=53); family practice teachers (19) beginning a seminar series on evidence based medicine.

Intervention The Fresno test is a performance based measure for use in medical education that assesses a wide range of evidence based medicine skills. Open ended questions are scored with standardised grading rubrics. Calculation skills are assessed by fill in the blank questions.

Main outcome measures Inter-rater reliability, internal reliability, item analyses, and construct validity.

Results Inter-rater correlations ranged from 0.76 to 0.98 for individual items. Cronbach's α was 0.88. Item difficulties ranged from moderate to difficult, all with positive and strong ability to discriminate between candidates. Experts scored consistently higher than novices. On the 212 point test, the novice mean was 95.6 and the expert mean was 147.5 ($P < 0.001$). On individual items, a higher proportion of experts than novices earned passing scores on 15 of the 17 items.

Conclusion The Fresno test is a reliable and valid test for detecting the effect of instruction in evidence based medicine. Its use in other settings requires further exploration.

Introduction

Medical educators need valid methods to assess instruction in evidence based medicine.[1] Existing tests assess subjective outcomes, such as attitude and self reported skill,[2] or only a single skill, such as critical appraisal.[3] The Fresno test of evidence based medicine was designed to assess the effectiveness of a comprehensive evidence based medicine curriculum in the University of California, San Francisco's Fresno family practice residency programme. The curriculum emphasises the process described by Sackett et al[4] with additional attention to the applicability or relevance of other recent discussions to your patient population.[5 6] The Fresno test assesses performance of each component of evidence based practice, rather than relying on self report. We describe the development, reliability, and validity of the test.

Methods

Description of test

The Fresno test begins with the presentation of two scenarios that suggest clinical uncertainty. Short answer questions about the clinical scenarios require the candidate to formulate a focused question, identify the most appropriate research design for answering the question, show knowledge of electronic database searching, identify issues important for determining the relevance and validity of a given research article, and discuss the magnitude and importance of research findings. These questions are scored by using a standardised grading system. A series of calculations and fill in the blank questions follow. The full questionniare is available on bmj.com.

Development of test

We wrote open ended test questions to reflect objectives of our course on evidence based medicine, beginning with formulation of a clinical question and continuing through critical appraisal of an article. Unlike multiple choice or true-false questions, the open ended questions require examinees to show higher order thinking in response to an authentic task.[7] The test concludes with calculations and fill in the blank questions that assess ability to apply some of the principles discussed in the short answer questions. We also developed scoring criteria based on predicted responses and our expert opinion about the elements of an ideal answer. To establish the face validity of the test, we distributed early drafts and grading rubrics to teachers of evidence based medicine. We removed controversial elements and adopted others in response to their suggestions.

We published the test on the world wide web and linked it to a database to store responses. Fresno University family practice residents and faculty members (n=43) took the test before formal instruction in evidence based medicine. In addition, 53 self identified experts, recruited through an email list server for evidence based medicine teachers, volunteered to take the test. No further measures of this

University of California San Francisco, Fresno Medical Education Program, Department of Family and Community Medicine, 445 South Cedar Avenue, Fresno, CA 93702, USA

Kathleen D Ramos *assistant professor*

Sean Schafer *associate professor*

California State University Fresno, Kremen School of Education and Human Development, 5005 North Maple Avenue, Fresno, CA 93740, USA

Susan M Tracz *professor*

Correspondence to: K D Ramos katie.ramos@ucsfresno.edu

BMJ 2003;326:319–21

P+

The test and further details of the validation process are available on bmj.com

Properties, measurements, and results of Fresno test

Test property	Measure used	Acceptable results	Performance of Fresno test
Content validity (test covers entire topic of interest)	Expert opinion	Test covers all the main aspects of evidence based medicine	Revisions based on experts' suggestions
Inter-rater reliability (degree to which 2 scorers rate a single performance similarly)	Inter-rater correlation (development dataset)	Expected to be high (⩾0.60)	Ranged from 0.76 to 0.98 for individual items, total scores 0.98
	Inter-rater correlation (validation dataset)	May be comparable to development set or slightly lower, but still >0.60	Similar to development dataset, ranged from 0.72 to 0.96 for individual items, total scores 0.97.
Internal reliability (degree to which all test questions on the test measure a single construct)	Cronbach's α—average of all possible split half correlations	⩾0.85	0.88
	Item-total correlation	⩾0.30	Ranged from 0.47 to 0.75 for individual items
Item difficulty (relative difficulty of each item)	% of candidates who answer achieve a passing score	Wide range of difficulties allows a test to be used with both expert and novice groups	Ranged from moderate (73%) to difficult (24%); no easy items
Item discrimination (ability of each item to discriminate between those with overall high scores and those with overall low scores)	Item discrimination index (ranges from –1.0 to 1.0)	All items should have a positive discrimination index, preferably >0.20	Ranged from 0.41 to 0.86, no items had negative or weak discriminations
Construct validity (evidence that the test measures the construct it intends to)	Mean scores of experts and novices compared by *t* test	Significant difference, higher expert scores	On a 212 point test, novice mean was 95.6 and expert mean was 147.5 (P<0.001)
	% passing for expert and novice groups compared by χ^2 test	Significant difference, higher % of experts passing	For all items a higher proportion of experts than novices passed; 15/17 differences were significant (P<0.05)

group's expertise were gathered. These 96 tests comprised the development data set.

We removed personal identifiers from the tests, and two of us (KR and SS) scored all of them independently. Through discussion, we revised the grading rubrics to minimise ambiguity and scored the tests again. After final revisions, we scored the tests a final time and calculated the test properties described below.

Because the iterative process described above for development of the grading rubrics could result in deceptively high inter-rater agreement for the development set, we used the final grading rubrics to score a validation set of new tests. These consisted of 19 tests taken by family practice teachers beginning a seminar series on evidence based medicine. For these tests, we report only inter-rater reliability to establish stability of this property.

Scoring the test

The raters scored the short answer questions using the standardised grading rubrics. For each question, the rubric specifies explicit grading criteria. For instance, the first item asks the respondent to write a focused clinical question. Responses are scored based on their inclusion of a patient population, an intervention, a comparison, and an outcome, each represented as a column on the rubric's table. The rows of the table represent four or five grading categories (not evident, minimal and/or limited, strong, excellent), each of which is associated with a point value. For instance, no mention of a patient population earns 0 points (not evident), the use of a general patient identifier is a limited answer (2 points), mentioning a single specific patient descriptor is a strong answer (4 points), and using numerous relevant descriptors is excellent (6 points). Each criterion is scored into these categories. The sum of points for all criteria is the score for that item.

For the item described above, limited performance in each category would result in a score of 8. We therefore considered any total less than 8 for a question as "not evident." A score of 8-15 was defined as a limited response, 16-23 as a strong response, and 24 as an excellent response. To assess the difficulty of items in the test we had to assign a cut off for a "passing" answer. We used our professional judgment of adequate mastery of the material[8] to set this cut off as the midpoint of the strong category of response. By this process, each short answer response is assigned a numerical score (from 0 to 24 points) and designated pass or fail. We scored calculations and fill in the blank questions as pass or fail and assigned points. The total test score is the sum of points for all items.

Results

The table describes the test properties in detail. Inter-rater reliability was excellent, both for the development dataset and for a validation dataset. Other test properties were derived solely from the development set of 96 tests. Internal consistency (Cronbach's α) was very good. Every item helps to distinguish candidates (item discrimination), and the items range in difficulty from moderate to difficult. Mean scores of the experts were significantly higher than the mean scores of the novices. More detail and discussion of validity and reliability is available on bmj.com.

Discussion

The Fresno test is a simple, reliable, and valid tool for assessing knowledge and skill in all the usual domains of evidence based medicine—asking focused questions, searching for good answers, critiquing literature, and applying conclusions in practice.[4] This comprehensive, performance based test has content validity, good to excellent inter-rater reliability for all questions, and excellent internal consistency. We established the stability of inter-rater reliability by validating a set of previously unscored tests. Construct validity, as measured by the ability of the test to distinguish between experts and novices, is high. Item difficulty is generally high but varies widely. No floor or ceiling effect was evident. This means that both novices and experienced practitioners can be assessed. However, the best use of the Fresno test is to measure change in knowledge after instruction in evidence based medi-

cine or to determine areas of weakness before instruction or practice.

The test may also be useful to show competency in evidence based medicine. A passing score could be defined by asking individuals who are agreed to meet or exceed minimum competence to take the test and setting a minimum proficiency score based on the range of these scores.

Validity and reliability

There are limitations to the validity, reliability, and general utility of the Fresno test. The groups we used to develop and validate the test probably represented the extremes of proficiency, leaving the middle ground relatively under-represented. The properties of the test may change when it used to assess groups of people that are more representative of the full range of proficiency in evidence based medicine.

The content of the test is based on the domains of evidence based medicine as promulgated by several widely read authors.[4 9] Nonetheless, there may be disagreement about whether these are the most relevant areas or about whether the questions and grading rubric accurately represent ideal content. For example, on the test item about external validity (or relevance) the expert group did not score significantly higher than the novice group. We chose to retain this item because it examines the recently emphasised issue of clinical relevance,[5 6] which we have found useful in our curriculum. As the evidence based medicine evolves, individual items may be more or less representative of current practice.

This test relies exclusively on the opinion of experts as the ultimate standard against which candidates are judged. Although expert opinion is the standard when developing tests, practising physicians are more concerned with improved patient outcomes. However, as no test exists that measures patient outcomes, the Fresno test is an improvement over current methods of assessing learning by self report.[1]

The inter-rater reliability reported here is high despite the inherent subjectivity of a test of this nature. The two raters participated in the construction and revision of the rubrics and therefore knew them well when scoring these tests. This familiarity with the rubrics may have led to unrealistically consistent scoring.

Also, the test presently has only one set of clinical vignettes and one set of numeric examples for calculation questions. We have written, but not tested, new clinical vignettes. Other vignettes will probably be needed if the test is used in other clinical disciplines.

Conclusions

The Fresno test is the first standardised, objective measure of ability in evidence based medicine that requires learners to demonstrate knowledge and skill. It can assess the effectiveness of teaching in evidence based medicine and identify strengths and weaknesses of curriculums and individuals. Further investigation might examine whether reliability and validity extends to new sets of raters and learners in other clinical disciplines and to other clinical vignettes. Medical educators may be further challenged to develop tests that reliably assess use of evidence based medicine in real clinical circumstances, not simulated or prompted by vignettes.

We thank John Smucney, Upstate Medical University, New York, for providing the validation data set.

What is already known on this topic

Instruction in evidence based medicine is provided in many medical education settings, but it effectiveness is unknown

Existing measures to assess competence tend to be narrowly focused and of uncertain validity

What this study adds

The Fresno test measures a wide range of knowledge and skills necessary for evidence based practice

The standardised grading systems produced a high degree of consistency between graders

Experts scored significantly higher on the test than novices in evidence based medicine, showing that the test has construct validity

Contributors: KDR participated in constructing the initial test and grading rubric, grading of test, and subsequent regrading of revisions, was responsible for revisions of rubric, administered test to subjects, solicited volunteer experts to take test, and wrote the manuscript. SS participated in constructing the initial test and grading rubric, grading of test, subsequent regrading of revisions, and writing the manuscript. SMT participated in constructing the initial grading rubric, statistical analysis of data, and wrote part of the methods section of the manuscript. KDR and SS are guarantors

Funding: This research was funded as part of a Residency Training in Primary Care grant to teach evidence based medicine, from the Bureau of Health Profession, Health Resources and Services Administration.

Competing interests: None declared.

1 Hatala R, Guyatt G. Evaluating the teaching of evidence-based medicine. *JAMA* 2002;288:1110-1.
2 Grad R, Macaulay AC, Warner M. Teaching evidence-based medical care: description and evaluation. *Fam Med* 2001;33:602-6.
3 Stern DT, Linzer M, O'Sullivan PS, Weld L. Evaluating medical residents' literature-appraisal skills. *Acad Med* 1995;70:152-4.
4 Sackett DL, Straus SE, Richardson WS, Rosenberg W, Haynes RB. *Evidence-based medicine: how to practice and teach EBM*. 2nd ed. London: Churchill Livingstone, 2000.
5 Slawson DC, Shaughnessy AF, Barry H. Which should come first: rigor or relevance? *J Fam Pract* 2001;50:209-10.
6 Slawson DC, Shaughnessy AF. Obtaining useful information from expert based sources. *BMJ* 1997;314:947-9.
7 Bloom BS. *Taxonomy of educational objectives, handbook I: the cognitive domain*. New York: McKay, 1956.
8 Wiersma W, Jurs S. *Educational measurement and testing*. 2nd ed. Boston: Allyn and Bacon, 1990.
9 Guyatt G, Rennie D, eds. *Users' guides to the medical literature: a manual for evidence-based clinical practice*. Chicago: American Medical Association Press, 2002.

(Accepted 16 October 2002)

Corrections and clarifications

Testing new pharmaceutical products in children

We inadvertently omitted to publish the name and affiliation details of the second author of this editorial (11 January, pp 64-5). We published only the details of Alastair G Sutcliffe, implying that he was the sole author; his coauthor, however, was Vic Larcher, a consultant paediatrician and paediatric ethicist at the Royal London Hospital, London EC1 2DP. We apologise for this error.

Involving patients can work in home blood glucose testing

The author of this letter, David Kerr, has informed us of authorship errors in reference 5 (11 January, pp 103-4). The authors are Ingleby J, Trowbridge S, Kerr D, Cavan DA.