Can Family Medicine Residents Predict Their Performance on the In-training Examination?

Robert W. Parker, MD; Cindy Alford, PhD; Cindy Passmore, MA

Background and Objectives: Evaluation of residents’ competence is of great importance in the training of physicians, yet the evaluation process is largely subjective. Faculty and residents frequently rely on self-assessments to make decisions regarding curriculum design and electives. The In-training Examination (ITE), the only widely available objective measure of residents’ medical knowledge, provides an opportunity to test the reliability of self-assessments. This study’s objective was to determine if family medicine residents are able to self-assess their medical knowledge by predicting their performance on the ITE.

Methods: A survey asking the residents to estimate their performance on the ITE in each of the nine content areas was administered at 13 examination sites just prior to the ITE. Correlation coefficients were calculated for corresponding predicted and actual scores for each resident in each content area. Predictions were also compared to performance according to quartile.

Results: Residents showed little ability to predict their scores in any of the content areas. Residents scoring in either the lowest or highest quartile were least able to predict accurately, with correct predictions ranging from 3% to 23%.

Conclusions: Residents cannot reliably predict their performance on the ITE. Of special concern are residents scoring in the lowest quartile, since these residents greatly overestimated their performance.

(Fam Med 2004;36(10):705-9.)

Medical educators currently emphasize self-directed learning for residents and physicians. An assumption that learners are aware of the strengths and weaknesses in their knowledge is implicit in such an approach. The In-training Examination (ITE), the only widely available objective measure of residents’ medical knowledge, provides an opportunity to test the reliability of such self-assessments. Research, using the ITE, cast doubt on the accuracy of residents’ self-assessment of their knowledge. A small study of 19 residents in a single family medicine program found a poor association between residents’ predicted scores on the ITE and their actual performance. Other studies show that physicians opt for continuing medical education (CME) in areas where they are already competent, rather than on topics in which they have knowledge deficits. Kruger and Dunning demonstrated in a series of studies how undergraduate students with deficiencies in knowledge had difficulty recognizing their own incompetence. In their studies, only students who were already competent were accurate in their self-assessments. In a similar study of self-assessment skills of medical students on a variety of cognitive tasks (such as interpreting electrocardiograms) and performance tasks (such as examination of a standardized patient), researchers found that the type of task does not affect self-assessment. In other words, self-assessment skills are generalized skills that operate similarly across a variety of tasks and contexts and are directly related to competence in the subject area.

Faculty predictions of residents’ performance on the ITE are also poor. More-experienced faculty were somewhat better at predicting residents’ scores, but generally, faculty were not reliable estimators of residents’ knowledge base as tested by the ITE.

Family medicine educators and the residents themselves need to know if they have good self-assessment skills, since self-assessment necessarily plays a key role in both teaching and learning. The question of whether the residents are aware of their areas of strength and weakness has not been adequately addressed in previous studies. This study’s purpose was to determine how effectively residents can assess their knowledge, specifically by predicting their scores on the different topic.

From the Department of Family Medicine, University of Texas Health Science Center at San Antonio (Drs Parker and Alford); and the Waco Faculty Development Center, Waco, Tex (Ms Passmore).
areas in the ITE. We also investigated whether knowledge of the subject matter was related to the ability to predict performance.

Methods
In November each year, family medicine residents take the ITE, which tests their knowledge in nine content areas. These content areas include clinical problem solving, gynecology, geriatrics, psychiatry, pediatrics, community medicine, obstetrics, surgery, and internal medicine. Internal medicine and geriatrics comprise 48% of the examination content.

Subjects
Requests for participation were sent to a convenience sample of 31 family medicine residency program directors in Texas and Oklahoma. Participation was voluntary. A total of 17 programs agreed to participate. Institutional Review Board approval was obtained by each residency program if it was required. Four programs reported incomplete or unusable data and were not included in the data analysis. The final study subjects thus consisted of 311 residents from 13 family medicine residency programs. Four of the 13 were university-based programs, and nine were community-based, university-affiliated programs. Gender distribution for the study population was 55% men and 45% women. The age range was 24 to 55 years, with a mean age of 30.9 years.

Design
A proctor at each location administered a one-page survey to each resident immediately before the ITE. Residents were asked to estimate how they would score in each of the nine ITE content areas. To ensure uniformity, the proctor at each site read a standard instructional script to the residents before they completed the survey. To ensure anonymity, each resident’s survey was assigned an identification code so that it could later be matched with his or her coded ITE result. Each institution kept the survey and examination codes of residents to ensure that the investigators would be blinded to the residents’ identities.

Survey Instrument
The survey instrument was a one-page document with simple instructions and a linear scale for self-rating of knowledge in each of the nine content areas on the ITE. The residents were asked to estimate their knowledge compared to other learners at their level of training in each of the nine ITE subject categories tested. They were to draw a vertical line on a 100 mm visual analog scale (VAS), which ranged from “weak” to “strong.” If a subject marks the midline on the VAS scale, the subject is predicting he or she will score better than 50% of all other residents taking the examination at their level. The mark on each VAS was subsequently measured and converted to numerical data (range 0 to 100) and paired with the result on ITE in the corresponding topic area, which was reported as a percentile rank within their postgraduate year (PGY) level at the national level (range 0% to 100%). Both exam scores and predicted scores were then categorized according to quartile for ease of comparison.

Data Analysis
The statistical software JMP was used to analyze the data. The Pearson correlation coefficients were calculated in each of the nine content categories for the paired pre-examination survey VAS and examination percentile rank. Scatter plots were examined for evidence of nonlinear correlation. We looked for patterns in performance based on gender and year of training. We next considered predictions and performance according to quartile, examining the data to see whether residents could predict the quartile of their performance. We were then able to consider whether performance on the test was related to the ability to predict accurately. Resident percentile scores for both the actual score and the predicted score were categorized according to quartile for each of the nine content areas.

Results
Residents’ predicted performance did not correlate strongly with their actual performance. Pearson correlation coefficients for all residents in each content area and for the subcategories of men and women were all less than .3. Results are shown in Table 1.

Table 2 shows the percent of residents who correctly predicted the quartile in which they would score in each of the nine subject areas. Success in predicting test scores according to quartile ranged from a low of 3% to a high of 54%. Those scoring in the middle two quartiles were better at predicting their performance than those scoring in either the top or the bottom quartiles. The ability of men and women to predict their ITE scores was similar (Table 1). Added years of training did not improve the residents’ ability to predict their performance.

The large majority of residents scoring in the bottom quartile overestimated their performance. Clinical problem solving was the content area in which poorly performing residents were most likely to overestimate their scores. Only 3% correctly predicted they would score in the lowest quartile. Since clinical problem solving is the educational goal, lack of self-knowledge on the part of residents is of great concern. Overall, residents correctly predicted their poor performance on the ITE only 10%–15% of the time. Figure 1 compares the number of residents who predicted they would score in
the lowest quartile to those who actually did score in the lowest quartile. Those scoring in the top quartile consistently underestimated their performance (Figure 2).

**Discussion**

The ability of residents to predict their performance in any of the ITE content areas was poor. The largest correlation coefficient was equal to .293 for men in the area of obstetrics, but, even with that degree of correlation, only 9% (r = .09) of the variation in the ITE scores can be explained by the variation in the predicted score. Therefore, the remaining 91% of the difference between predicted and actual ITE scores must be explained by other factors, such as motivation, test-taking ability, or impaired self-assessment.

The residents scoring in the lowest quartile on the examination are of special concern, since they need remediation the most but generally do not recognize that their knowledge is poor. For example, in the internal medicine content area, 84 of 302 residents scored in the lowest quartile, but only nine predicted their performance in that lowest quartile (Figure 1). This means that residents most in need of internal medicine knowledge improvement are unaware of their weakness and would not know they need remediation. In the same manner, those who scored in the highest quartile were examined (Figure 2). Only 14 of the 67 residents scoring in the top quartile in internal medicine correctly predicted their performance. However, underestimating performance carries little risk and may even be a positive sign. This underestimation of knowledge may even serve as a stimulus for further study among the most successful residents. Indeed, acting on a perceived need for additional study may be the reason why these residents score highly.

Our results are of a similar magnitude to those from a small study of 19 residents in a single family medicine program.\(^5\) When residents participating in that study were asked to predict their scores on the ITE, a low association between residents’ predicted ITE scores and the actual ITE score (r = .21) was found. Our findings also support the conclusions of Kruger and Dun-

### Table 1

<table>
<thead>
<tr>
<th>Percentile Correlation Coefficients by Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Clinical</strong></td>
</tr>
<tr>
<td><strong>All entries are r</strong>&lt;br&gt;(P Value)</td>
</tr>
<tr>
<td>All</td>
</tr>
<tr>
<td>Women</td>
</tr>
<tr>
<td>Men</td>
</tr>
</tbody>
</table>

* Test for null hypothesis r = 0, significant at P ≤ .05

### Table 2

<table>
<thead>
<tr>
<th>Percent of Residents Who Accurately Predicted Their Test Score, by Quartile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Area</td>
</tr>
<tr>
<td>Internal medicine</td>
</tr>
<tr>
<td>Surgery</td>
</tr>
<tr>
<td>Pediatrics</td>
</tr>
<tr>
<td>Psychiatry</td>
</tr>
<tr>
<td>Obstetrics</td>
</tr>
<tr>
<td>Community medicine</td>
</tr>
<tr>
<td>Geriatrics</td>
</tr>
<tr>
<td>Gynecology</td>
</tr>
<tr>
<td>Clinical problem solving</td>
</tr>
</tbody>
</table>

n=302
Thus, residents who are performing poorly are the least likely to be aware of their deficiencies. This is an important finding, since faculty often rely on resident input to choose topics for study. Curricula are designed around learner needs, which are often determined by resident surveys and interviews. Educators’ presentations and the content are evaluated by residents, which may lead to the elimination of the topic or the presentation. Another concern is that residents may choose electives by personal preference rather than by learning needs. We believe faculty should oversee curriculum choices and develop an individualized plan for each poorly performing resident.

Limitations
This study has several limitations. First, the subjects were limited to a specific geographic area. There is no reason, however, to believe that residency programs and residents in Texas and Oklahoma are substantially different from those in the rest of the country. Second, the American Board of Family Practice specifically states that the ITE was not designed to evaluate individuals. The ITE is a norm-referenced test and is intended to evaluate residency programs, not individual residents. It has little validity as a measure of individual performance. Nevertheless, program directors and faculty commonly use the examination results in the nine content areas to evaluate residents in need of remediation and to modify their residency program’s curriculum content.

Conclusions
The results of this study indicate that underperforming residents do not have a good assessment of their medical knowledge, ie, they have poor self-assessment skills. Self-assessment is invaluable in becoming a lifelong learner. It is important since residency programs are required by the Accreditation Council for Graduate Medical Education to train residents to become lifelong learners in practice-based learning, one of the six basic competencies.

Does the same inability to identify strengths and weaknesses affect practicing physicians? Survey techniques similar to those used in our study could answer this question, using the American Board of Family Practice recertification examination instead of the ITE. Such a study would determine if further clinical experience improves self-assessment.

Acknowledgments: We thank E. Mikaila Adams for her editorial and technical assistance in the preparation of this manuscript.

Financial support: This publication was supported by Grant No. 5D45PE56003 from the Division of Medicine of the Health Resources and Services Administration (HRSA). Its contents are solely the responsibility of the authors and do not necessarily represent the views of HRSA. This material is provided in part through a grant from the Texas Higher Education Coordinating Board.

Corresponding Author: Address correspondence to Dr Alford, University of Texas Health Science Center at San Antonio, 7703 Floyd Curl Drive, MC 7795, San Antonio, TX 78229-3900. 210-358-3200. Fax: 210-223-6940. alford@uthscsa.edu.
REFERENCES


