Comparing Diagnostic Performance and the Utility of Clinical Vignette-Based Assessment Under Testing Conditions Designed to Encourage Either Automatic or Analytic Thought

Jonathan S. Ilgen, MD, MCR, Judith L. Bowen, MD, Lucas A. McIntyre, Kenny V. Banh, MD, David Barnes, MD, Wendy C. Coates, MD, Jeffrey Druck, MD, Megan L. Fix, MD, Diane Rimple, MD, Lalena M. Yarris, MD, MCR, and Kevin W. Eva, PhD

Abstract

**Purpose**
Although decades of research have yielded considerable insight into physicians’ clinical reasoning processes, assessing these processes remains challenging; thus, the authors sought to compare diagnostic performance and the utility of clinical vignette-based assessment under testing conditions designed to encourage either automatic or analytic thought.

**Method**
This 2011–2012 multicenter randomized study of 393 clinicians (medical students, postgraduate trainees, and faculty) measured diagnostic accuracy on clinical vignettes under two conditions: one encouraged participants to give their first impression (FI), and the other led participants through a directed search (DS) for the correct diagnosis. The authors compared accuracy, feasibility, reliability, and relation to United States Medical Licensing Exam (USMLE) scores under each condition.

**Results**
A 2 (instructional condition) × 2 (vignette complexity) × 3 (experience level) analysis of variance revealed no difference in accuracy as a function of the correct diagnosis. The authors compared accuracy, feasibility, reliability, and relation to United States Medical Licensing Exam (USMLE) scores under each condition.

**Conclusions**
Instructions to trust one’s FI result in performance in a testing environment under the FI condition relative to the DS condition.

The diagnostic reasoning literature suggests that expert performance does not stem from the application of generic problem-solving strategies. Rather, when tackling a clinical problem, individuals appear to link the information they gather to their own idiosyncratic memories, to sort rapidly between possible diagnoses, and concurrently to identify data that might not “fit” their hypotheses. Models codifying these complex processes have recently centered on two systems of cognition, termed System 1 and System 2. System 1, which likely drives most first impressions, describes the automatic, rapid, subconscious processes that are activated when individuals encounter familiar problems; System 2, which is more deliberate, describes the analytic, slow, conscious processes that individuals employ when considering novel or ambiguous problems. Although distinguished as explanatory models, the cognitive processes used in actual clinical reasoning typically involve both systems, an interaction referred to as “dual processing.” Using both systems in concert may offer the best chance at diagnostic success. In an effort to further this translation, we hypothesized that we could gain unique insights by giving instructions in an assessment setting that biased participants toward particular modes of reasoning. We expect that many aspects of clinical practice (e.g., time pressures and distractions) push clinicians towards System 1 processing. Thus, it is plausible that testing conditions that similarly promote System 1 processing might more accurately reflect performance in practice. At the same time, if the deliberate application of System 2 processing is essential for improved diagnostic performance, then testing conditions that emphasize automatic approaches could underestimate candidates’ actual ability. Knowing whether deliberate instructions to use a particular diagnostic approach do or do not improve performance in a testing environment can help advance assessment practices. Further, providing explicit directions to use a specific diagnostic approach also...
affords a direct test of whether cognitive debiasing strategies aimed at promoting System 2 processing help clinicians avoid System 1–induced errors.15,16 Answering such questions cleanly is difficult because System 1 and System 2 rarely, if ever, operate alone.20 Using manipulative instructions and altering the conditions in which clinicians make decisions can, however, influence their diagnostic accuracy; such strategies have been used to demonstrate that both automatic and analytic processing provide value as well as detriment.21

Whether or not such manipulations can influence the utility of diagnostic assessment strategies remains unknown; therefore, we developed an online vignette-based assessment that guides participants toward using either automatic or analytic reasoning processes to make a diagnosis.22 Previously, we tested participants with different levels of clinical experience at a single institution, and our results suggested that straightforward vignettes under both testing conditions provided similar reliability for medical students, whereas complex vignettes under instructions that encouraged automatic thinking offered the greatest reliability for PGY1 and PGY2 residents.22 Since then, research has shown that faster responses within a test correlate with better performance when clinicians are encouraged to “work as quickly as [they] can without sacrificing accuracy”; this finding further suggests that diagnostic accuracy is a function of individuals’ baseline knowledge and experience with specific problems more than the application of any particular decision-making processes.23

The current investigation, which has two aims, extends this research. First, to determine whether instructions that encourage clinicians to use more deliberate thinking improve diagnostic accuracy, we compared clinicians’ performance under conditions designed to encourage either automatic or analytic reasoning processes. Second, to evaluate the impact of these instructions on the utility of the assessment protocol, we examined the reliability of the results, the feasibility of the testing administration, and the results’ correlation with United States Medical Licensing Exam (USMLE) scores under both conditions. Accordingly, we sought to answer four questions:

1. Does diagnostic accuracy improve when clinicians are encouraged to consider clinical vignettes through deliberate processes (as opposed to automatic processes)?
2. Does the ability of the assessment to consistently differentiate among candidates (i.e., reliability) differ across testing conditions?
3. Does the relationship between performance on the assessment and scores on high-stakes licensing examinations differ across testing conditions?
4. When balancing testing time and reliability, how feasible is it to administer each testing condition?

**Method**

**Design**

This multicenter, randomized, cross-sectional study measured clinicians’ diagnostic accuracy on a series of previously developed clinical vignettes under two instructional conditions designed to encourage either automatic or analytic reasoning. We constructed our clinical assessment using a Web-based testing platform. The institutional review board at the coordinating institution approved this study. As compensation for their time, participants were entered into a raffle to win one of three Apple iPads (Cupertino, California).

**Participants**

We recruited third- and fourth-year medical students, along with residents, fellows, and faculty from emergency and internal medicine from eight academic institutions in the western United States between December 2011 and January 2012. We grouped participants into three cohorts based on their clinical experience: students, residents/fellows, and faculty.

**Procedure**

We invited eligible participants (total N = 3,757) via e-mail during a two-week enrollment period at each site. The e-mail they received included a brief description of the study and a link to the testing Web site. On clicking the link, respondents were directed to an open-source blog that provided more information about the study, including that the investigators would collaborate with the National Board of Medical Examiners (NBME) to link performance on the vignette-based diagnostic assessment to past scores on the USMLE. Through the Web site, we collected participants’ consent and demographic information. Further, we programmed the Web site such that when participants clicked the button corresponding to their training level, it automatically randomized each of them into one of six assessment groups (three per instructional condition) with stratification by training level to ensure an even distribution across the six versions of the assessment (see Figure 1). The NBME confidentially linked assessment results with participants’ USMLE scores at the conclusion of the study and returned an anonymized dataset to the investigators for analysis.

We designed each instructional condition to encourage participants to use either automatic or analytic reasoning processes to solve clinical vignettes. All of the instructions for participants in both conditions appeared on a single computer screen. To encourage automatic reasoning, we drew from the work of Ark and colleagues.14 Half of the participants received testing instructions encouraging them to use similarity as their guide and to record their first impressions (FI) as their diagnosis:

> Please read the vignette below. This case may look similar to things you have seen before (i.e., during training): trust your sense of familiarity and enter [in the box that followed] the first diagnosis that seems reasonable upon reading the case.

To encourage analytic forms of reasoning, we used an abbreviated set of prompts similar to those employed by Mamede and colleagues.25 Thus, the remaining half of participants received instructions designed to facilitate a directed search (DS) for the correct diagnosis; their instructions comprised four prompts asking them to explain their reasoning:

1. Think about how you would put together the information you have just been given about this case into a one-sentence summary. Write that one-sentence summary in the box below, using the terms you feel most accurately and concisely sum up the case.

2. List each diagnosis you are considering. For each diagnosis, write the features of the history, exam, and/or test results that support that diagnosis, and features that do not explain that diagnosis.

3. Now list no more than three most probable diagnoses for this case in rank
Participants recruited by e-mail (N = 3,757) with link to study site
- 8 academic medical centers
- 3rd- and 4th-year medical students
- Emergency medicine and internal medicine residents, fellows, and faculty

Participants (N = 393) provided informed consent. Participants identified their level of training, then the computer randomized them into, first, one of two testing conditions and, then, within that condition, one of three assessment versions:

- First impression instructions (N = 201)
  - Exam A (N = 64)
  - Exam B (N = 68)
  - Exam C (N = 67)

- Directed search instructions (N = 192)
  - Exam A (N = 67)
  - Exam B (N = 65)
  - Exam C (N = 60)

Figure 1 Study flow diagram, illustrating the allocation of participants taking an eight-item, vignette-based diagnostic assessment to one of six different exam versions, 2011–2012. The authors collected data only from participants who completed the entirety of their assessments. Exam A contained vignettes 1 through 8; Exam B contained vignettes 5 through 12; Exam C contained vignettes 1 through 4 and 9 through 12.

Each version of the assessment began with a training vignette that demonstrated the expected free-text answer format for each instructional condition, followed by eight testing vignettes. A previous publication describes the instrument development and pilot process. Briefly, we selected cases from a pool of clinical vignettes developed and tested by Mamede and colleagues. Vignettes reported patients’ demographic and historical information, physical examination findings, and lab or other test results. We used 12 vignettes covering a spectrum of complexity: half were “simple” vignettes that described features of a typical presentation of a single diagnosis, and half were “complex” vignettes that intentionally introduced features that suggested more than one diagnosis. We previously published validity evidence for these vignettes in the domains of content, response process, relation to other variables, and internal structure.

We delivered the assessment to participants via Catalyst (Catalyst Development Corporation, Yucca Valley, California), an online software program that allows for asynchronous testing. We divided the 12 vignettes into three blocks of four, each with two “simple” cases and two “complex” cases. We constructed three versions of the assessment by combining each of the three blocks in a pairwise fashion such that each assessment included eight vignettes (see Figure 1). We instructed participants not to use supplemental material, not to discuss the vignettes with others, and to allot sufficient time to complete the assessment in one uninterrupted sitting. The Catalyst software automatically measured total testing times.

Measurements
Using a previously described rubric, two raters (J.S.I. and J.L.B.) scored all responses; generalizability coefficients were then calculated to evaluate the interrater reliability. Under the FI condition, we allowed only a single diagnosis; if a participant entered more than one diagnosis in the free-text box, we scored the entire response as incorrect. For comparability, we scored responses in the DS condition as correct only if a participant had ranked the correct diagnosis first. For discrepant scores, we came to a consensus through discussion before assigning a score for analysis. The interrater reliability of our scoring was high (G > 0.99) for each assessment.

Analysis
We measured and compared diagnostic accuracy and overall time required across instructional conditions using analysis of variance (ANOVA). We also compared results of the assessment under the two conditions with USMLE scores using Pearson correlation coefficients.

Because the main outcome of interest was reliability achieved under each instructional condition (FI and DS), we based our sample size calculations on the equations presented by Streiner and Norman. To calculate reliability meaningfully, yet separately for simple and complex cases, we assumed the use of four vignettes and a reliability of at least 0.75. We calculated that we would require a sample of N = 53 per instructional condition to achieve a 95% confidence interval (CI) of ≥ 0.05. To consider the reliability separately for each experience level, we aimed to enroll 60 participants × 6 groups (2 conditions × 3 experience levels) for a total N = 360. Pilot testing revealed that a sample this size should provide > 95% power to detect a 10% difference in performance between groups with moderate (d = 0.5) effect size.

We used generalizability theory to calculate the internal consistency of the assessment scores for each level of experience. This allowed us to analyze how well our diagnostic reasoning assessment could consistently differentiate individuals within a group of peers, without artificially inflating the reliability by analyzing across the full (and, hence, more heterogeneous) sample. Generalizability theory also enabled us to account for the three replications of the effect of test ( instructional) condition. We treated participants as the facet of differentiation and vignette as the facet of generalization. We then performed a series of Spearman–Brown calculations to examine assessment feasibility by determining how many questions are required within each condition to achieve reliability of at least 0.75. To examine the feasibility of testing within each condition, we multiplied the proportional increase in the number of items by the observed testing time.

Results
Figure 1 illustrates the trial flow, and Table 1 provides the characteristics of the study participants. A total of 393 individuals participated in the study, though we excluded 5 who provided nonsensical or partial responses.
Table 1
Characteristics of Participants Completing an Eight-Item Vignette-Based Diagnostic Assessment, Winter 2011–2012

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No. (% of 393) participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Institution*</td>
<td></td>
</tr>
<tr>
<td>Oregon Health &amp; Science University</td>
<td>24 (6)</td>
</tr>
<tr>
<td>UC Davis</td>
<td>32 (8)</td>
</tr>
<tr>
<td>UC Los Angeles</td>
<td>45 (11)</td>
</tr>
<tr>
<td>UC San Francisco</td>
<td>48 (12)</td>
</tr>
<tr>
<td>University of Colorado</td>
<td>22 (6)</td>
</tr>
<tr>
<td>University of New Mexico</td>
<td>30 (8)</td>
</tr>
<tr>
<td>University of Utah</td>
<td>17 (4)</td>
</tr>
<tr>
<td>University of Washington</td>
<td>175 (45)</td>
</tr>
<tr>
<td>Training level</td>
<td></td>
</tr>
<tr>
<td>Medical students</td>
<td>157 (40)</td>
</tr>
<tr>
<td>Residents/fellows</td>
<td>97 (25)</td>
</tr>
<tr>
<td>Faculty</td>
<td>139 (35)</td>
</tr>
<tr>
<td>Specialty*</td>
<td></td>
</tr>
<tr>
<td>Emergency medicine</td>
<td>90 (23)</td>
</tr>
<tr>
<td>Internal medicine</td>
<td>146 (37)</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>199 (51)</td>
</tr>
<tr>
<td>Female</td>
<td>194 (49)</td>
</tr>
</tbody>
</table>

*UC indicates University of California.

1This number of participants in this category excludes the 157 students and, hence, the percentages do not sum to 100.

Testing time

Overall mean testing time was 74.9 minutes (standard deviation [SD] = 149.8). Of the 393 participants we included, 6 (1.5%) had testing times classified as outliers because they were > 3 SDs from the mean. Excluding these individuals, mean testing times in the FI and DS conditions were 29.1 minutes (95% CI = 23.8–34.5) and 88.6 minutes (95% CI = 83.4–93.9), respectively (F[1,381] = 242, P < .001, d = 0.03), nor did discipline and condition interact (F[1,381] < 1, P = .60). To conserve statistical power, we did not consider the discipline of participants in any further analyses.

Table 2 shows diagnostic accuracy as a function of instructional condition, vignette complexity, and experience level. A 2 (instructional condition) × 2 (complexity) × 3 (experience level) ANOVA revealed no difference as a function of instructional condition (F[1,228] = 2.44, P = .12, d = 0.12), but demonstrated the expected main effects of complexity (F[1,228] = 965.2, P < .001, d = 1.92) and experience (F[2,228] = 39.6, P < .001, d = 0.73). We found no statistically significant interactions among these three variables apart from experience × complexity (F[2,228] = 6.17, P = .002), which revealed that the difference in accuracy between simple and complex vignettes declined with experience, although the effect size was large in all instances. Mean differences were 48.6% (d = 2.11), 46.2% (d = 2.01), and 37.6% (d = 1.63) for, respectively, medical students, residents/fellows, and faculty.

Correlates with other performance measures

Table 3 presents Pearson correlations examining the relationship between average diagnostic assessment scores and USMLE performance, birth year, and testing time. Despite including only eight vignettes, we observed statistically significant correlations with each step of the USMLE. This trend was true both when we considered all data (i.e., derived from FI and DS instructions together) and when we analyzed data collected from the two conditions separately. In 11 out of 11 comparisons between instructional conditions (shown in the first three rows of Table 3), the FI instruction yielded a higher correlation between our diagnostic assessment scores and USMLE scores than did the DS instruction (P < .05). We also observed a small negative correlation between overall performance and birth year in both instructional conditions such that older (more experienced) participants performed slightly better on average.

Performance on the assessments did not correlate significantly with testing time in any instance.

Assessment reliability by instructional condition

To assess the internal consistency of our diagnostic assessment, we calculated Cronbach alpha for each of the three assessment versions. Table 4 shows the average reliability as a function of complexity and instructional condition. Table 4 also presents the results of Spearman–Brown calculations performed with these coefficients to determine the number of items (N(i)) required within each experimental condition to achieve reliability of at least 0.75 and the time requirements to complete an assessment of that length. These analyses revealed that an alpha of at least 0.75 can be achieved much more efficiently in the FI condition relative to the DS condition, although testing times remain long in both conditions.

Discussion

Results from this large, experientially diverse cohort of clinicians suggest that diagnostic performance is similar when participants are encouraged to trust their first impressions compared with when they are encouraged to take a more deliberate approach. Further, guiding clinicians into an automatic reasoning mode seems to have improved both the efficiency of the diagnostic assessment and the correlation of the assessment’s results with performance scores on the USMLE (Table 3)—despite the number of vignettes remaining constant. These results suggest that instructions can have a substantial impact on the ability of vignette-based diagnostic assessments to differentiate meaningfully among clinicians for reasons independent of the gains that can be achieved through increased sampling.

This large-scale implementation of our assessment intervention attempted to address some of the limitations of our initial single-institution study.22 The lack of difference in mean diagnostic accuracy between instructional conditions stands in contrast with our prior results in which diagnostic accuracy was greater in the DS condition.22 These differences likely stem from methodological differences between the two studies. In our first study, we presented the FI and DS conditions in sequence, such that...
Table 2
Mean Diagnostic Accuracy of 393 Participants Completing an Eight-Item, Vignette-Based Diagnostic Assessment, 2011–2012 as a Function of the Three Independent Variables: Instructional Condition, Vignette Complexity, and Subject Experience

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Influence of each independent variable on diagnostic accuracy</th>
<th>Mean accuracy measured as % correct (95% confidence intervals)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Instructional condition</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First impression</td>
<td>F(1,379) = 2.44, P = .12, d = 0.12</td>
<td>63.8 (61.4–66.3)</td>
</tr>
<tr>
<td>Directed search</td>
<td></td>
<td>61.1 (58.7–63.5)</td>
</tr>
<tr>
<td><strong>Vignette complexity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simple</td>
<td>F(1,379) = 965.2, P &lt; .001, d = 1.92</td>
<td>84.6 (82.5–86.6)</td>
</tr>
<tr>
<td>Complex</td>
<td></td>
<td>40.4 (38.1–42.8)</td>
</tr>
<tr>
<td><strong>Experience level</strong></td>
<td>F(2,379) = 39.6, P &lt; .001, d ≤ 0.73*</td>
<td></td>
</tr>
<tr>
<td>Medical students</td>
<td></td>
<td>62.7 (60.0–65.3)</td>
</tr>
<tr>
<td>Residents/fellows</td>
<td></td>
<td>65.2 (61.8–68.6)</td>
</tr>
<tr>
<td>Faculty</td>
<td></td>
<td>69.6 (66.8–72.4)</td>
</tr>
</tbody>
</table>

* Students versus faculty d = 0.73; students versus residents d = 0.54; residents versus faculty d = 0.19.

Table 3
Pearson Correlations Examining the Relationship Between Average Scores on Vignette-Based Diagnostic Assessment and United States Medical Licensing Exam Performance, Birth Year, and Amount of Testing Time as a Function of Instructional Condition of 393 Participants Completing the Eight-Item Diagnostic Assessment, 2011–2012

<table>
<thead>
<tr>
<th>Item</th>
<th>Medical students</th>
<th>Residents/fellows</th>
<th>Faculty</th>
<th>All groups combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>First impressions</td>
<td>Directed search impressions</td>
<td>First impressions</td>
<td>Directed search impressions</td>
</tr>
<tr>
<td></td>
<td>score*</td>
<td></td>
<td>score*</td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td>0.45*</td>
<td>0.29*</td>
<td>0.18</td>
<td>0.13</td>
</tr>
<tr>
<td>Step 2</td>
<td>0.36*</td>
<td>0.35*</td>
<td>0.40*</td>
<td>0.03</td>
</tr>
<tr>
<td>Step 3</td>
<td>0.41*</td>
<td>0.14</td>
<td>0.39*</td>
<td>0.28</td>
</tr>
<tr>
<td>Birth year</td>
<td>0.04</td>
<td>0.14</td>
<td>0.08*</td>
<td>0.03</td>
</tr>
<tr>
<td>Testing time</td>
<td>0.14</td>
<td>0.16</td>
<td>0.07</td>
<td>0.04</td>
</tr>
</tbody>
</table>

* Step 1, 2, and 3 scores refer to the United States Medical Licensing Examination. Medical student participants had not yet taken the Step 3 exam.
* P < .05.
robustness of context specificity in diagnostic reasoning. Extrapolation of our data suggests that increasing the number of items would achieve decent reliability, but the already-long testing times associated with the DS condition likely preclude feasible implementation of that intervention. The finding that testing under the FI condition resulted in consistently higher correlations with USMLE performance than did the DS condition with a constant number of vignettes suggests that either less time can be spent on the vignette-based diagnostic assessment or that time can be equated (i.e., using longer testing times, but encountering more clinical vignettes) to enable better content coverage without harming other aspects of the assessment’s utility or candidates’ performance.

Although internal consistency was relatively low in all groups (Table 4), a few interesting observations should be highlighted. If anything, DS instructions better differentiated between medical students relative to FI instructions, whereas the reverse was true (if anything) for more experienced clinicians. Again, this finding is not surprising given that truly difficult vignettes are too difficult for even the best trainees, and truly easy vignettes are too easy for even the weakest experienced diagnosticians.

Our study has several limitations. First, we used volunteer participants, which may have resulted in a sample with more competent or motivated individuals. This influence, however, should narrow the range of scores and result in more conservative reliability estimates. Further, such a bias would be unlikely to affect the results systematically across the instructional conditions given the random assignment of participants. Second, we collected data only when participants completed their entire assessment. Possibly then, because of the substantially more laborious nature of the DS condition, more of the participants randomized to that condition started but did not finish their assessment. Strengths of this investigation include the use of realistic clinical vignettes that have been vetted in prior investigations, the large sample size that is experientially diverse, and the deliberate collection of validity evidence to ground these results in the context of existing literature and past performance.

In conclusion, this investigation suggests that instructions to trust one’s first impressions result in similar diagnostic accuracy when compared with instructions to consider clinical information in a structured and systematic fashion. These results run counter to the assertion that encouraging clinicians to focus their thinking toward analytic processes is likely to reduce diagnostic errors. Further, instructions to trust first impressions created a substantial advantage in terms of the utility of the assessment protocol, enabling more feasible test administration in a manner that appears to correlate better with high-stakes licensing examinations.

---

### Table 4

Psychometric Properties of an Eight-Item, Vignette-Based Diagnostic Assessment

<table>
<thead>
<tr>
<th>Group</th>
<th>First impression</th>
<th>Directed search</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Observed alpha</td>
<td>To achieve alpha ≥ 0.75</td>
</tr>
<tr>
<td></td>
<td>(8 items)</td>
<td>N(i)¹</td>
</tr>
<tr>
<td>All participants</td>
<td>0.42</td>
<td>34</td>
</tr>
<tr>
<td>Students</td>
<td>0.36</td>
<td>53</td>
</tr>
<tr>
<td>Residents</td>
<td>0.25</td>
<td>133</td>
</tr>
<tr>
<td>Faculty</td>
<td>0.32</td>
<td>72</td>
</tr>
</tbody>
</table>

*Alpha indicates internal consistency of test scores averaged over eight items.

¹N(i) indicates the number of items required to achieve alpha ≥ 0.75.

²Time indicates the number of hours required to complete a test of length N(i).

---

---

6

---

Acknowledgments: The authors thank Silvia Mamede, MD, PhD, for providing the clinical vignettes and initial translation for this study. They acknowledge the assistance of Erika Goldstein, MD, MPH, and Molly Jackson, MD, for enrolling subjects at the University of Washington.

Funding/Support: This work was supported by a grant from the Association of American Medical Colleges Western Group on Educational Affairs (WGEA).

Other disclosures: None.

Ethical approval: The University of Washington School of Medicine institutional review board approved this study.

---

**Dr. Ilgen** is assistant professor, Division of Emergency Medicine, University of Washington, School of Medicine, Seattle, Washington.

**Dr. Bowen** is professor, Department of Medicine, Oregon Health & Science University, School of Medicine, Portland, Oregon.

**Mr. McIntyre** is a student, University of Washington, School of Medicine, Seattle, Washington.

**Dr. Banh** is assistant clinical professor, Division of Emergency Medicine, University of California, San Francisco, School of Medicine, Fresno, California.

**Dr. Barnes** is assistant professor, Department of Emergency Medicine, University of California, Davis, School of Medicine, Sacramento, California.

**Dr. Coates** is professor of clinical medicine, Department of Emergency Medicine, University of California, Los Angeles, David Geffen School of Medicine, and Los Angeles Biomedical Research Institute, Los Angeles, California.

**Dr. Druck** is associate professor, Department of Emergency Medicine, University of Colorado, School of Medicine, Denver, Colorado.

**Dr. Fix** is assistant professor, Division of Emergency Medicine, University of Utah, School of Medicine, Salt Lake City, Utah.

**Dr. Rimple** is associate professor, Department of Emergency Medicine, University of New Mexico, School of Medicine, Albuquerque, New Mexico.

**Dr. Druck** is associate professor, Department of Emergency Medicine, Oregon Health & Science University, School of Medicine, Portland, Oregon.

**Dr. Yarris** is professor and director of education research and scholarship, Department of Medicine, and senior scientist, Centre for Health Education Scholarship, University of British Columbia, Vancouver, British Columbia, Canada.
References