Published in final edited form as:
Acad Emerg Med. 2011 October ; 18(Suppl 2): S79–S86. doi:10.1111/j.1553-2712.2011.01182.x.

Adjusting Our Lens: Can Developmental Differences in Diagnostic Reasoning Be Harnessed to Improve Health Professional and Trainee Assessment?

Jonathan S. Ilgen, MD, MCR, **Judith L. Bowen, MD**, **Lalena M. Yarris, MD, MCR**, **Rongwei Fu, PhD**, **Robert A. Lowe, MD, MPH**, and **Kevin Eva, PhD**
Division of Emergency Medicine, University of Washington School of Medicine, (JSI) Seattle, WA; Department of Medical Informatics and Clinical Epidemiology (JLB), Department of Emergency Medicine (LMY, RF, RAL), Oregon Health & Science University, Portland, OR Centre for Health Education Scholarship, University of British Columbia, (KE) Vancouver, BC

Abstract

Objectives—Research in cognition has yielded considerable understanding of the diagnostic reasoning process and its evolution during clinical training. This study sought to determine whether or not this literature could be used to improve the assessment of trainees' diagnostic skill by manipulating testing conditions that encourage different modes of reasoning.

Methods—The authors developed an online, vignette-based instrument with two sets of testing instructions. The "first impression" condition encouraged non-analytic responses while the "directed search" condition prompted structured analytic responses. Subjects encountered six cases under the first impression condition, then six cases under the directed search condition. Each condition had three straightforward (simple), and three ambiguous (complex) cases. Subjects were stratified by clinical experience: novice (third and fourth year medical students), intermediate (post graduate year [PGY] 1 and 2 residents), and experienced (PGY 3 residents and faculty). Two investigators scored the exams independently. Mean diagnostic accuracies were calculated for each group. Differences in diagnostic accuracy and reliability of the examination as a function of the predictor variables were assessed.

Results—The exam was completed by 115 subjects. Diagnostic accuracy was significantly associated with the independent variables of case complexity, clinical experience, and testing condition. Overall, mean diagnostic accuracy and the extent to which the test consistently discriminated between subjects (i.e., yielded reliable scores) was higher when participants were given directed search instructions than when they were given first impression instructions. In addition, the pattern of reliability was found to depend on experience: simple cases offered the best reliability for discriminating between novices, complex cases offered the best reliability for discriminating between intermediate residents, and neither type of case discriminated well between experienced practitioners.

Conclusions—These results yield concrete guidance regarding test construction for the purpose of diagnostic skill assessment. The instruction strategy and complexity of cases selected should

Corresponding Author: Jonathan S. Ilgen, MD, MCR, University of Washington School of Medicine, Division of Emergency Medicine, Harborview Medical Center, 325 9th Ave, Box #359702, Seattle, WA 98104-2499, ilgen@u.washington.edu.

Prior Presentations: Presented at the 17th Annual National Research Service Award (NRSA) Trainees Research Conference. June 2011, Seattle WA.

Funding Sources/Disclosures: This publication was made possible with support from the Oregon Clinical and Translational Research Institute (OCTRI), grant number UL1 RR024140 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. The authors have no conflicts of interest to report.

depend on the experience level and breadth of experience of the subjects one is attempting to assess.

INTRODUCTION

The Institute of Medicine (IOM) report, *To Err Is Human: Building a Safer Health System*, revealed a staggering burden of medical errors in American health care.[1] Hospital-wide studies of these adverse events demonstrated a particularly high proportion of diagnostic errors in the emergency department (ED).[2,3] While these events often stem from multiple factors, 96% of missed diagnoses in the ED have been linked to cognitive errors.[4] Past work suggests that these types of errors arise from faulty interpretation, synthesis, and judgment, rather than insufficient data gathering or fund of knowledge,[4–7] and experts from a diversity of fields have offered insights for how to more effectively teach these cognitive skills.[8–16] Untested to this point is whether or not the lessons learned from this literature can be harnessed to improve health professional and trainee assessment.

The research reported here addresses this measurement gap. Current theories of cognition suggest that clinicians use different problem-solving strategies depending on their clinical experience, familiarity with the problem at hand, and case ambiguity.[17–20] Diagnosticians at all levels employ non-analytic reasoning when they encounter familiar problems.[10,13,21,22] These subconscious processes, such as pattern recognition or heuristics, allow clinicians to recognize clinical presentations and quickly link them to past experiences.[23] Alternatively, cases that are novel or ambiguous tend to trigger analytic reasoning, a slower, more effortful process where clinicians consciously sort through information to prioritize hypotheses. Research has shown that accuracy improves when novices are prompted to reason through clinical problems using a combined (analytic and non-analytic) approach.[24] Developmentally, we expect novice clinicians to rely more on analytic reasoning approaches, while more experienced clinicians have a larger mental database of cases from which to match clinical patterns via non-analytic reasoning.[9, 25–29]

We have developed an online instrument to assess clinicians' diagnostic accuracy in the context of two unique sets of testing instructions, each designed to bias subjects toward either analytic or non-analytic reasoning. We hypothesized that these testing conditions could provide stratification of individuals at different developmental stages of diagnostic reasoning and, in addition, would provide different degrees of reliability across experience level. Here we report on these findings.

METHODS

Study Design

This was a cross-sectional study. The institutional review board at Oregon Health & Science University approved this project. Written informed consent was obtained from all participants.

Study Setting and Population

We recruited third- and fourth-year medical students, and residents and faculty from the departments of emergency medicine (EM) and internal medicine (IM), from an urban academic medical center. Participants were divided into three cohorts based on their clinical experience: "novice" (medical students), "intermediate" (first- and second-year residents), and "experienced" (third-year residents and faculty). All eligible students, residents, and faculty were invited to participate via e-mail during a two-week period of enrollment. Participants received $5 coffee shop gift cards as compensation for their time, and were entered into a raffle to win an Apple iPad. Consent and demographic information were obtained during the online test. We were unable to capture data regarding non-respondents through our initial e-mail inquiry.

Study Protocol

The online examination consisted of 12 vignettes, divided evenly between two conditions with unique sets of testing instructions. These instructions were designed to bias subjects toward one of two modes of reasoning (analytic versus non-analytic). Subjects entered responses using free-text boxes. Examinations were scored with a rubric, and we calculated subjects' diagnostic accuracy.

Testing authorities in medical education currently advocate for the use of authentic vignette-based clinically oriented questions.[30,31] We initially selected 20 clinical vignettes from a collection of cases previously developed and tested by Mamede et al.[17,18] These cases reported patients' demographic and historical information, physical examination findings, and test results. We selected cases with a spectrum of ambiguity: "simple" vignettes described features of a typical presentation of a single diagnosis, while "complex" vignettes intentionally introduced features that suggested more than one diagnosis.[18] Performance comparisons across these conditions and across different levels of experience serve as a test of validity of the test scores generated by the testing protocol. Because these vignettes were initially written in Portuguese and tested in Brazil, we made slight adjustments to the English language translation and incorporated normal lab value ranges for reference. These cases were built into an online instrument using Sakai (Sakai Foundation, Ann Arbor, MI), an open-source software package that allows for asynchronous testing. The exam was then pilot-tested by 14 faculty members, seven from each discipline. These subjects' scores and responses were not included in the study dataset. Based on feedback, the instrument was shortened to 12 vignettes, with six simple and six complex cases. A think-aloud protocol with six additional faculty was then conducted to ensure agreement of acceptable free-text answers for each case, and a scoring rubric was created.

Using the pseudo-random number generator function of SAS (SAS Institute, Cary, NC), we distributed the 12 cases into unique locations in four exams. Placement of these cases was stratified such that subjects encountered an equal number of simple and complex cases during the first half and second half of the examination, respectively. The four exams were counterbalanced such that the same cases were presented in both testing instructions at equal rates.

Instructions for the first half of the exam were aimed at encouraging a non-analytic approach to diagnosing the clinical problem by instructing participants to diagnose each case by offering their "first impression." Instructions for the second half of the exam were aimed at encouraging an analytic approach to diagnosing the clinical problem by explicitly directing participants to reason through the case in a particular way ("directed search"; see Figure 1). To familiarize the subjects with the testing instructions and answer format, a training vignette preceded each condition. Following the final vignette, we collected demographic information including sex, age, discipline, and year in training. We then asked for feedback from the participants, specifically whether they were able to locate a suitable testing station at the time they took the test, whether they found the testing website easy to navigate, whether the instructions and questions were written clearly, and whether the length of the test seemed reasonable and appropriate. The Sakai software automatically measured testing times as the difference between when subjects began the first vignette and completed the twelfth vignette.

We tested subjects during a 10-day period in May, 2010. At the time of testing, third-year medical students had gained exposure to all of their core rotations. In an effort to obtain spontaneous, unbiased responses to the cases, participants were instructed to not use any supplemental material and to not discuss the cases with others. Two investigators scored the examinations using the predefined rubric. Under the first impression condition, only a single response was generated, and was scored as correct or incorrect. If a subject listed more than one diagnosis in the free-text box under the first impression condition, it was scored as incorrect. Responses under the directed search condition were scored as correct only if the correct diagnosis was ranked first among the possible diagnoses named. When there was disagreement between the two investigators' scoring, consensus was obtained through discussion before assigning a score for final analysis.

Measures

The primary predictor variable was the type of instruction received, first impression or directed search. Secondary independent variables included the complexity of the clinical vignettes, level of clinical experience, and medical specialty. The outcome of interest was diagnostic accuracy.

Given that the standard deviation (SD) for similar instructional conditions was previously demonstrated to be ±0.25,[17] and a moderate to large effect size of 0.5 has been shown to reflect judgments of clinical importance,[32] sample size calculations indicated a requirement of 64 participants per condition.

Data Analysis

Analyses were conducted using SAS version 9.2, PASW Statistics version 18.0 (SPSS Inc., IBM, Armonk, NY), and Microsoft Excel (Microsoft Corp., Redmond, WA). Demographics were summarized using descriptive statistics. Kappa statistics were calculated to evaluate inter-rater agreement for free-text responses. To assess whether subjects' diagnostic accuracy improved as a function of growing accustomed to the structured answer format (i.e. a "learning effect" or "test order" effect), we examined the diagnostic accuracy of simple and complex vignettes versus their ordinal position in the sequence of vignettes and tested the time trend using a linear regression model after adjusting for instruction type. To assess the effect of each independent variable on diagnostic accuracy, accuracy was submitted as a dependent variable to a 2 (type of instruction: first impression or directed search) × 2 (vignette difficulty: simple or complex), × 3 (training level: novice, intermediate, or experienced) analysis of variance (ANOVA). These independent variables entered the model as fixed effects. All participants' mean scores were confirmed to lie within three standard deviations from the mean, and Geisser-Greenhouse and Huynh-Feldt corrections were performed to ensure that any heterogeneity of variance did not affect the conclusions. To assess the effect of the test instructions and vignette difficulty on the psychometric properties of the test, a one-way ANOVA using subject as the grouping factor and vignette as a nested variable (given that the condition to which each vignette was assigned varied by subject) was conducted to enable variance components to be extracted and intraclass correlation coefficients (ICCs) to be calculated. The ICCs reflect the extent to which the assessment offered reliable data (i.e., consistently discriminated between the subjects). These coefficients were calculated for the dataset as a whole, as well as separately for each condition within and across training levels. Given that reliability is affected by the number of vignettes scored, decision study analyses were conducted to equate the reliability observed within each experimental condition with the reliability that would be anticipated were a 12-vignette exam presented using that testing format alone. Ninety-five percent confidence intervals (CI) were calculated for each ICC to enable comparisons.

RESULTS

Of 444 eligible subjects, 158 (35.5%) agreed by e-mail to participate, and 115 of those (72.3%) completed the test, for a total response rate of 26%. Participant characteristics are shown in Table 1. Of these respondents, 98.2% reported being able to find a suitable testing station, 95.4% claimed they could navigate the site easily, 88.1% indicated the questions were clearly written, and 65.1% thought the test was a reasonable length. On average, the test required 1.17 hours to complete. When scores from two investigators were compared, there was disagreement for only 7 of the 1,470 answers (Cohen's kappa 0.99; 95% CI = 0.985 to 0.998).

Mean diagnostic accuracy as a function of instruction type, degree of difficulty, and training level is reported in Table 2. No significant differences in diagnostic accuracy were detected for specialty, age, or sex. These were not shown to be important confounding variables, so these results are not illustrated. The linear trend after adjusting for type of instruction was not significant ($p = 0.96$ for simple cases, and $p = 0.94$ for complex cases). ANOVA revealed the anticipated main effects of experience level (greater experience being associated with higher performance; $F(2,112) = 22.4$, mean squared error [MSE] = 0.08, $p < 0.001$) and vignette complexity (complex cases being associated with poorer accuracy; $F(1,112) = 368.3$, MSE = 19.1, $p < 0.001$). Overall, accuracy was higher in the directed search condition compared to the first impression condition ($F(1,112) = 9.1$, MSE = 0.6, $p < 0.01$). None of the two-way or three-way interactions were statistically significant. Post-hoc analyses revealed that the mean scores of novice subjects were significantly lower than those of both intermediate and experienced subjects ($p < 0.001$). There was no significant difference between the mean scores of experienced and those of intermediate subjects ($p = 0.18$).

Reliability analyses performed on the entire set of 12 vignettes revealed an ICC(1,12) of 0.35 (95% CI = 0.32 to 0.39). Thus, 35% of the variance in scores was attributable to the test takers themselves, with the remainder being better attributed to case differences, a case-by-test taker interaction, and unaccounted-for sources of error. Decision studies revealed that a 52-item test of this type would be required to achieve an ICC of 0.70 for a similarly heterogeneous sample.

Intriguing differences emerged when the vignettes were analyzed separately to assess the effect of test instruction and case complexity on the reliability. Table 3 illustrates the reliabilities observed as a function of both item complexity and test instruction. More straightforward (i.e., simple) cases generally yielded greater reliability than did complex cases. Similarly, the directed search instructions generally led to a test with greater reliability than did the first impression instructions, despite the vignettes being the same in both experimental conditions.

As reliability indicates the extent to which one can consistently discriminate between subjects, we expected that the heterogeneity of our subjects' clinical experience would affect these measurements. We thus performed similar analyses using subjects from each level of experience independently. Interesting variations from the above patterns of reliability were observed. Regardless of test instructions, the scores received by novice participants were greater for simple cases (ICC(1,12) = 0.59 and 0.61 for first impression and directed search cases, respectively) than for complex cases (ICC(1,12) = 0.00). In contrast, the scores received by intermediate participants were greater for complex cases (ICC(1,12) = 0.59 and 0.33 for first impression and directed search cases, respectively) than for simple cases (ICC(1,12) = 0.00). Finally, neither simple nor complex cases provided reliable discrimination of experienced candidates (ICC(1,12) = 0.00 in all assessments except for simple cases presented in the context of first impression instructions, ICC(1,12) = 0.23).

DISCUSSION

Considerable research has sought to define and understand clinical reasoning, demonstrating that both analytic and non-analytic processes exist and can facilitate diagnostic accuracy.[7,21,24,29,33–38] The instrument and pilot results reported in this study provide innovative progress in this domain by measuring accuracy and reliability in the context of instructions that influence how subjects solve clinical problems. Because we expect certain reasoning strategies to be dominant at different stages of experience, we anticipated that this approach would highlight valuable cognitive differences among trainees.

Our data suggest that this online instrument is feasible for asynchronous assessment of a spectrum of learners and can be reliably scored. Prior studies of diagnostic reasoning that used vignettes and methods similar to ours required proctored exams and paper-based data collection.[17–19] Although easily navigated and amenable to remote administration, testing times and subjective feedback from participants suggest completion rates could improve if the test took less time to complete. Future studies will gauge whether reliability can be maintained or improved within an acceptable amount of testing time by simply employing one of the two instructional conditions rather than both in combination.

We observed that instructions given to participants influenced the psychometric properties of test scores from our vignette-based instrument. When the entire continuum of experience, from novices to experienced practitioners, was considered in a single analysis, greater reliability was observed in the directed search instructional condition than in the first impression condition. This suggests that more experienced practitioners are better differentiated from their novice counterparts through their ability to deliberately work through a case, whereas first impressions differentiate novices from experts less well. That the instrument was not able to discriminate between experienced subjects suggests that once a certain diagnostic reasoning skill level has been reached, practicing physicians are largely indistinguishable from one another when presented with clinical vignettes of the type presented in this study. Conversely, replicable differences did exist in intermediates, although only when complex cases were presented, and in novices only when simple cases were presented. The lack of difference in reliability across test instructions for novice diagnosticians suggests that replicable performance differences between individuals at this level of training exist both in the capacity to quickly recognize the correct diagnosis, and the capacity to deliberately reason through a case presentation. When complex cases were presented to the intermediate experience group, replicable differences existed only when participants were instructed to adopt a more non-analytic (i.e., first impression) mode of reasoning. This is consistent with the notion that performance in medicine is less related to overall knowledge and more related to an ability to apply that knowledge in a judicious and accurate way.[22,24,29,39,40]

It is important to recognize that no instructional intervention can create a pure comparison between different reasoning processes. It is equally conceivable that subjects instructed to rely on their first impression also deliberately considered features, and that subjects given deliberate search instructions were influenced by their first impressions. With our instructional manipulation we only hope to have prompted differential weighting of the two reasoning approaches. In particular, because the deliberate search instructions followed the first impression instructions, the directed search instructional condition may have performed as a “combined” reasoning approach. Previous work has demonstrated equivalent performance between subjects who received simultaneous or sequential instructions to use combined reasoning, and the diagnostic performance of these subjects exceeded that for subjects who received only one set of instructions (either first impression or directed search).[24]

Past work by Mamede and colleagues has investigated the reasoning processes of second-year Brazilian IM residents.[17–19] Using similar testing instructions and vignettes, but employing a repeated measures design that alternated instructional condition from one case to the next, these researchers failed to demonstrated differences in accuracy between the two instructional conditions.[17] The significantly higher diagnostic accuracy under the directed search condition that we demonstrated in this study could be due to the experiential heterogeneity of our subject population, or subjects' use of a sequential combined reasoning approach for the second block of six questions, for which they received deliberate search instructions.

This is the first study to apply these clinical vignettes to a population with heterogeneous clinical experience. The significant associations between mean scores of diagnostic accuracy and both clinical experience and vignette complexity suggest that the scores provide a valid reflection of clinical skill. Further study is planned to assess how individuals' scores change longitudinally with training, and the degree to which scores from this instrument correlate with external markers of clinical performance. If our instrument is used to follow a cohort of learners over time, results may provide useful information about the development of clinical reasoning skills. If applied to distinguish between cohorts, it would appear that this instrument is most useful to distinguish between "novice" students and more experienced residents and faculty. Within-group comparisons are likely to provide more meaningful and useful information about performance differences between experiential peers. Further testing with a larger multicenter population will determine whether meaningful differences in performance persist between subjects within each respective cohort.

LIMITATIONS

The sequential nature with which the instructions were presented is the main limitation of this study. It is possible that the reliability of the directed search condition was better than that of the first impression condition simply because those instructions were presented last, thereby introducing a greater amount of error into the measurement in the first half of the test. However, our results align with the existing clinical reasoning literature. Further, the consistent differences in reliability seen as functions of vignette complexity and experience level were not afflicted by a similar confound. Nonetheless, this issue remains to be tested in future research.

Second, we used volunteer subjects, which may have resulted in a sample that was more confident, able, or motivated than a typical population of medical students, residents, and faculty. However, this should narrow the range of scores, thus yielding conservative reliability estimates. Furthermore, it is unlikely that such a bias in the sample would create a systematic differential in the reliability observed across experimental conditions.

Third, our subjects were drawn from a single institution. Since medical schools and academic health centers have variable prevalence of disease pathologies, the experiences of these subjects with the clinical material used in the vignettes could differ from those of individuals at another institution. It is also possible that some training institutions have specific training in diagnostic reasoning itself. These concerns may limit our ability generalize these findings to other sites.

Fourth, the cases employed in this testing instrument were drawn from a collection that was originally developed for an audience of Brazilian IM residents.[17–19] While this may raise concerns regarding the validity of this content to the subjects tested in this study, we were careful to choose cases with pathologies that our subjects would routinely encounter.

Finally, though mean scores under each testing condition were higher among subjects with more clinical experience, this study did not compare these scores to clinical evaluations provided by supervisors, scores on other instruments designed to evaluate reasoning,[41–44] or reasoning documented in targeted case reviews.[4,5,45] This limits interpretation of whether these scores are a valid reflection of reasoning in a true clinical environment, as well as whether this instrument adds value to other available tools.

CONCLUSIONS

Our online vignette-based testing instrument was feasible for administration to an experientially diverse subject group, and could be scored reliably. Differences in diagnostic accuracy followed expected patterns by experience levels and vignette complexity, supporting the argument that these scores are a valid reflection of reasoning performance. Reliability studies yielded intriguing results, suggesting instructional conditions influence the psychometric properties of the instrument. Decisions regarding whether to use simple or complex vignettes and which test instructions to use should be driven by the level and breadth of the experience within which test administrators hope to differentiate.

Acknowledgments

The authors thank Silvia Mamede for providing the clinical vignettes and initial translation for this study. We thank John Ansorge for helping to build the Sakai testing platform. The authors thank the internal medicine and emergency medicine faculty members who generously volunteered their time to test the instrument and provide feedback.

References

1. Kohn, LT.; Corrigan, J.; Donaldson, MS. Institute of Medicine (U.S.). To Err is Human: Building a Safer Health System. Washington, DC: National Academies Press; 2000. Committee on Quality of Health Care in America.
2. Brennan TA, Leape LL, Laird NM, et al. Incidence of adverse events and negligence in hospitalized patients: results of the Harvard Medical Practice Study I. N Engl J Med. 1991; 324(6):370–6. [PubMed: 1987460]
3. Leape LL, Brennan TA, Laird N, et al. The nature of adverse events in hospitalized patients: results of the Harvard Medical Practice Study II. N Engl J Med. 1991; 324(6):377–84. [PubMed: 1824793]
4. Kachalia A, Gandhi TK, Puopolo AL, et al. Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. Ann Emerg Med. 2007; 49(2):196–205. [PubMed: 16997424]
5. Singh H, Thomas EJ, Petersen LA, Studdert DM. Medical errors involving trainees: a study of closed malpractice claims from 5 insurers. Arch Intern Med. 2007; 167(19):2030–6. [PubMed: 17954795]
6. Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med. 2005; 165(13):1493–9. [PubMed: 16009864]
7. Graber M. Diagnostic errors in medicine: a case of neglect. Jt Comm J Qual Patient Saf. 2005; 31(2):106–13. [PubMed: 15791770]
8. Bowen JL. Educational strategies to promote clinical diagnostic reasoning. N Engl J Med. 2006; 355(21):2217–25. [PubMed: 17124019]
9. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2005; 39(1):98–106. [PubMed: 15612906]
10. Kassirer JP. Teaching clinical reasoning: case-based and coached. Acad Med. 2010; 85(7):1118–24. [PubMed: 20603909]

11. Round AP. Teaching clinical reasoning--a preliminary controlled study. Med Educ. 1999; 33(7): 480–3. [PubMed: 10354329]
12. Goss JR. Teaching clinical reasoning to second-year medical students. Acad Med. 1996; 71(4): 349–52. [PubMed: 8645397]
13. Eva KW, Hatala RM, Leblanc VR, Brooks LR. Teaching from the clinical reasoning literature: combined reasoning strategies help novice diagnosticians overcome misleading information. Med Educ. 2007; 41(12):1152–8. [PubMed: 18045367]
14. Graber ML. Educational strategies to reduce diagnostic error: can you teach this stuff? Adv Health Sci Educ Theory Pract. 2009; 14(Suppl 1):63–9. [PubMed: 19669922]
15. Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med. 2003; 78(8):775–80. [PubMed: 12915363]
16. Croskerry P, Wears RL, Binder LS. Setting the educational agenda and curriculum for error prevention in emergency medicine. Acad Emerg Med. 2000; 7(11):1194–200. [PubMed: 11073467]
17. Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Med Educ. 2008; 42(5):468–75. [PubMed: 18412886]
18. Mamede S, Schmidt HG, Rikers RM, Penaforte JC, Coelho-Filho JM. Breaking down automaticity: case ambiguity and the shift to reflective approaches in clinical reasoning. Med Educ. 2007; 41(12):1185–92. [PubMed: 18045371]
19. Mamede S, Schmidt HG, Rikers RM, Penaforte JC, Coelho-Filho JM. Influence of perceived difficulty of cases on physicians’ diagnostic reasoning. Acad Med. 2008; 83(12):1210–6. [PubMed: 19202502]
20. Norman GR, Eva KW. Doggie diagnosis, diagnostic success and diagnostic reasoning strategies: an alternative view. Med Educ. 2003; 37(8):676–7. [PubMed: 12895245]
21. Croskerry P. A universal model of diagnostic reasoning. Acad Med. 2009; 84(8):1022–8. [PubMed: 19638766]
22. Norman GR, Brooks LR. The non-analytical basis of clinical reasoning. Adv Health Sci Educ Theory Pract. 1997; 2(2):173–84. [PubMed: 12386407]
23. Eva KW, Norman GR. Heuristics and biases--a biased perspective on clinical reasoning. Med Educ. 2005; 39(9):870–2. [PubMed: 16150023]
24. Ark TK, Brooks LR, Eva KW. Giving learners the best of both worlds: do clinical teachers need to guard against teaching pattern recognition to novices? Acad Med. 2006; 81(4):405–9. [PubMed: 16565197]
25. Eva KW. The aging physician: changes in cognitive processing and their impact on medical practice. Acad Med. 2002; 77(10 Suppl):S1–6. [PubMed: 12377689]
26. Eva KW, Norman GR, Neville AJ, Wood TJ, Brooks LR. Expert-novice differences in memory: a reformulation. Teach Learn Med. 2002; 14(4):257–63. [PubMed: 12395489]
27. Schmidt HG, Boshuizen HPA. On acquiring expertise in medicine. Educat Psychology Rev. 1993; 5(3):205–21.
28. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med. 1990; 65(10):611–21. [PubMed: 2261032]
29. Schmidt HG, Rikers RM. How expertise develops in medicine: knowledge encapsulation and illness script formation. Med Educ. 2007; 41(12):1133–9. [PubMed: 18004989]
30. The National Board of Medical Examiners (NBME). [Accessed Jul 28, 2011] Item Writing Manual. Available at: http://www.nbme.org/publications/item-writing-manual-download.html
31. Case SM, Swanson DB, Becker DF. Verbosity, window dressing, and red herrings: do they make a better test item? Acad Med. 1996; 71(10 Suppl):S28–30.
32. Norman GR, Sloan JA, Wyrwich KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. Med Care. 2003; 41(5):582–92. [PubMed: 12719681]
33. Croskerry P. The cognitive imperative: thinking about how we think. Acad Emerg Med. 2000; 7:1223–31. [PubMed: 11073470]

34. Norman G. Research in clinical reasoning: past history and current trends. Med Educ. 2005; 39(4): 418–27. [PubMed: 15813765]

35. Norman G. Dual processing and diagnostic errors. Adv Health Sci Educ Theory Pract. 2009; 14(Suppl 1):37–49. [PubMed: 19669921]

36. Norman G, Young M, Brooks L. Non-analytical models of clinical reasoning: the role of experience. Med Educ. 2007; 41(12):1140–5. [PubMed: 18004990]

37. Norman GR, Eva KW. Diagnostic error and clinical reasoning. Med Educ. 2010; 44(1):94–100. [PubMed: 20078760]

38. Graber M, Gordon R, Franklin N. Reducing diagnostic errors in medicine: what's the goal? Acad Med. 2002; 77(10):981–92. [PubMed: 12377672]

39. Norman GR, Brooks LR, Colle CL, Hatala RM. The benefit of diagnostic hypotheses in clinical reasoning: experimental study of an instructional intervention for forward and backward reasoning. Cognit Instruct. 1999; 17(4):433–48.

40. Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. Acad Med. 2000; 75(2):182–90. [PubMed: 10693854]

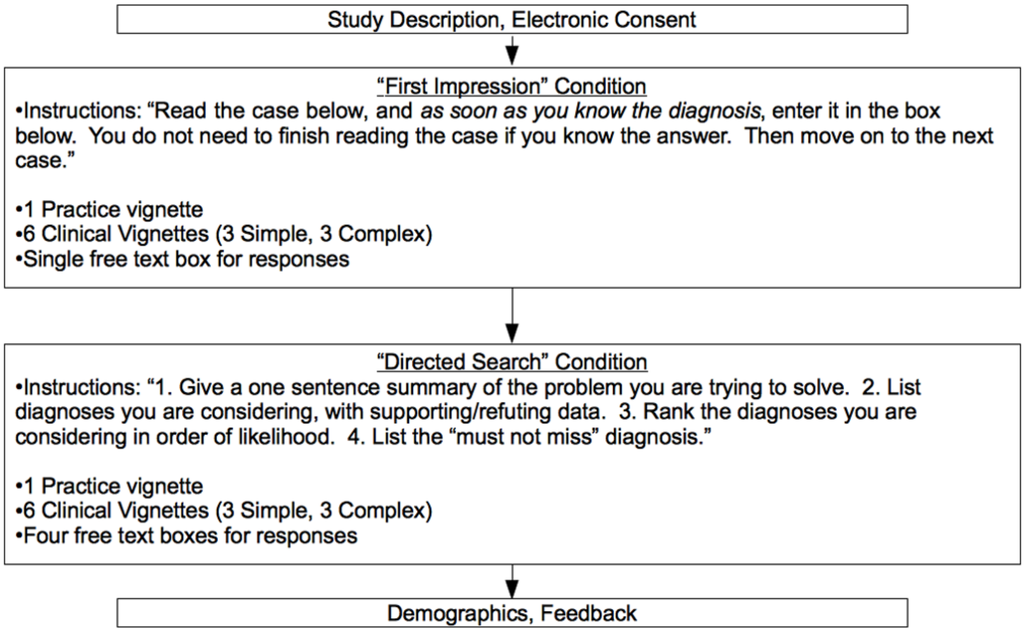
41. Brailovsky C, Charlin B, Beausoleil S, Cote S, Van der Vleuten C. Measurement of clinical reflective capacity early in training as a predictor of clinical reasoning performance at the end of residency: an experimental study on the script concordance test. Med Educ. 2001; 35(5):430–6. [PubMed: 11328512]

42. Cook DA, Erwin PJ, Triola MM. Computerized virtual patients in health professions education: a systematic review and meta-analysis. Acad Med. 2010; 85(10):1589–602. [PubMed: 20703150]

43. Dillon GF, Clauser BE. Computer-delivered patient simulations in the United States Medical Licensing Examination (USMLE). Simul Healthc. 2009; 4(1):30–4. [PubMed: 19212248]

44. Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. Med Educ. 2005; 39(12):1188–94. [PubMed: 16313577]

45. Chellis M, Olson J, Augustine J, Hamilton G. Evaluation of missed diagnoses for patients admitted from the emergency department. Acad Emerg Med. 2001; 8:125–30. [PubMed: 11157287]



Study Description, Electronic Consent

“First Impression” Condition

•Instructions: “Read the case below, and *as soon as you know the diagnosis*, enter it in the box below. You do not need to finish reading the case if you know the answer. Then move on to the next case.”

•1 Practice vignette
•6 Clinical Vignettes (3 Simple, 3 Complex)
•Single free text box for responses

“Directed Search” Condition

•Instructions: “1. Give a one sentence summary of the problem you are trying to solve. 2. List diagnoses you are considering, with supporting/refuting data. 3. Rank the diagnoses you are considering in order of likelihood. 4. List the “must not miss” diagnosis.”

•1 Practice vignette
•6 Clinical Vignettes (3 Simple, 3 Complex)
•Four free text boxes for responses

Demographics, Feedback

Figure 1.

Test administration flow diagram

Table 1

Demographics of study participants.

Cohort	Eligible for Enrollment	Agreed to Participate (%)	Completed Test (%)	Mean Age (SD)	Female (%)
Novice					
MS3	113	36 (31)	26 (20)	29.1 (5.0)	65.4
MS4	133	36 (27)	25 (19)	29.5 (3.9)	56.0
Intermediate					
PGY1	47	17 (36)	14 (30)	29.3 (2.9)	50.00
PGY2	39	15 (38)	12 (31)	30.3 (4.1)	91.7
Experienced					
PGY3	43	23 (53)	16 (37)	30.9 (2.0)	56.3
Faculty	69	31 (44)	22 (32)	44.5 (7.9)	36.4

MS: medical student; PGY: Post-graduate year.

Table 2

Diagnostic accuracy by testing condition and degree of difficulty for three cohorts of subjects.

			Instruction			
			First Impression		Directed Search	
Difficulty	Cohort	N	Mean Score (%)	95% CI	Mean Score (%)	95% CI
Simple	Novice	51	2.1 (69.9)	2.0–2.2	2.4 (79.7)	2.3–2.5
	Intermediate	26	2.4 (80.8)	2.2–2.6	2.7 (89.7)	2.5–2.9
	Experienced	38	2.6 (89.5)	2.5–2.9	2.7 (89.5)	2.5–2.8
Complex	Novice	51	0.7 (24.2)	0.6–0.9	0.9 (31.4)	0.8–1.1
	Intermediate	26	1.2 (39.7)	1.0–1.4	1.4 (47.4)	1.2–1.6
	Experienced	38	1.4 (47.4)	1.3–1.6	1.7 (55.3)	1.5–1.8

Table 3

Reliability* (and 95% confidence intervals) observed as a function of case complexity and test instructions.

	Case complexity		
Test instruction	**Simple vignettes**	**Complex vignettes**	**Row total**
First impression	0.57 (0.54–0.60)	0.00 (0.00–0.00)	0.08 (0.06–0.10)
Directed search	0.78 (0.76–0.80)	0.18 (0.15–0.21)	0.44 (0.40–0.48)
Column total	0.52 (0.48–0.56)	0.37 (0.34–0.41)	0.35 (0.32–0.39)

*The marginal values are not simple averages of the reliabilities in the cells. Rather, each cell corresponds to the reliability observed for the three cases that fit within that complexity/test instruction combination and the marginal corresponds to the reliability observed for the six cases that fit within the relevant row/column. In each instance a decision study was used to equate the reliability to what could be expected had all 12 vignettes used in the study been of the type described. Each, then, corresponds with a Case 1 ICC averaged across 12 vignettes (ICC(1,12)) according to Cohen and Fleiss' nomenclature.