Blink or Think: Can Further Reflection Improve Initial Diagnostic Impressions?

Brian J. Hess, PhD, Rebecca S. Lipner, PhD, Valerie Thompson, PhD, Eric S. Holmboe, MD, and Mark L. Graber, MD

Abstract

Purpose

Experienced clinicians derive many diagnoses intuitively, because most new problems they see closely resemble problems they've seen before. The majority of these diagnoses, but not all, will be correct. This study determined whether further reflection regarding initial diagnoses improves diagnostic accuracy during a high-stakes board exam, a model for studying clinical decision making.

Method

Keystroke response data were used from 500 residents who took the 2010 American Board of Internal Medicine (ABIM) Internal Medicine Certification Examination. Data included time to initial response on each question, whether the answer was correct, and whether or not the resident changed her or his initial response. The focus was on 80 diagnosis questions that comprised realistic clinical vignettes with multiple-choice single-best answers. Cognitive skill (ability) was measured using overall exam scores. Case complexity was determined using item difficulty (proportion of examinees that correctly answered the question). A hierarchical generalized linear model was used to assess the relationship between time spent on initial responses and the probability of correctly answering the questions.

Results

On average, residents changed their responses on 12% of all diagnosis questions (or 9.6 questions out of 80). Changing an answer from incorrect to correct was almost twice as likely as changing an answer from correct to incorrect. The relationship between response time and accuracy was complex.

Conclusions

Further reflection appears to be beneficial to diagnostic accuracy, especially for more complex cases.

Substantial evidence suggests that 10% to 15% of medical diagnoses are wrong or egregiously delayed.[1] Only a small fraction of these errors result in serious harm, but hundreds of thousands of diagnoses are made every day, such that the aggregate number of errors is staggering.[2] According to one estimate, diagnostic error accounts for 40,000 to 80,000 deaths annually in the United States.[3]

Cognitive shortcomings can be identified in two-thirds of all cases of diagnostic error.[4] This occasionally reflects insufficient knowledge or faulty data gathering, but the most common failure is in clinical reasoning and not synthesizing all of the available information. According to the dual-process paradigm of "how doctors think," many clinical problems are immediately recognized, and the diagnosis is made instantaneously and intuitively. In cases where the problem is not recognized, the diagnosis must be made by deliberate, rational analysis, a slower and laborious process.[5] The dual-process paradigm provides a convenient framework to discuss clinical reasoning, but it oversimplifies the complexity of the process by focusing on the two extremes of this "cognitive continuum."[6,7]

Experts and experienced clinicians rely heavily on intuition because so many new problems closely resemble problems they've seen before, and the vast majority of diagnoses reached via intuition will be correct. Just as in our everyday lives, intuitive decisions and responses are so often correct that we inherently trust them.[8] Intuition, however, is predictably fallible, and the same is true of intuitively derived diagnoses.[9,10] The reliability of intuitive diagnoses can be degraded by a host of factors known to affect intuitive processing, such as a human tendency toward premature closure (failing to consider other possibilities once a diagnosis is identified), cognitive load, bias induced by the case context, or our emotions.[11,12] If intuition is sometimes error prone, then should physicians trust their intuition or systematically reconsider their initial impressions?

One approach recommended by educators and cognitive scientists is to practice reflectively.[13–15] The literature suggests that clinical reasoning can be improved by querying an initial diagnostic impression,[16] "slowing down when you should,"[17] or taking a "diagnostic time out."[18] This would encourage clinicians to pause before ordering tests or treatments and deliberately reconsider the diagnosis. Mamede et al[14] and Ilgen et al[19] have shown that reflective consideration improves diagnostic accuracy of trainees under experimental conditions. Chess players, both expert and nonexpert, also improve performance with added deliberation for both simple and difficult challenges.[20]

Although these studies demonstrate that reflection is advantageous in controlled experiments, Ilgen and colleagues[21] have also reported contrary results under different experimental instructions—deliberate consideration did not improve performance compared with one's first impression. In summary, the experimental evidence on the value of deliberate consideration is mixed, perhaps because of different experimental conditions. To address this problem, we took advantage of a natural experiment

Please see the end of this article for information about the authors.

Correspondence should be addressed to Dr. Hess, Hess Consulting, 272 Rue du Replat, St-Nicolas, Québec, G7A 5E4 Canada; telephone: (418) 496-1958; e-mail: brianhessconsulting@gmail.com.

Acad Med. 2015;90:112–118.
First published online November 4, 2014
doi: 10.1097/ACM.0000000000000550

and studied the impact of both response time and changing initial diagnostic impressions on diagnostic accuracy using a high-stakes board exam with realistic clinical scenarios. Because it is difficult to study the effects of further reflection on clinical reasoning in actual practice, this approach provides insight in clinical information processing, while avoiding the confounding issues that arise in experimental studies.

Method

Instruments

We used data from the 2010 American Board of Internal Medicine (ABIM) Internal Medicine Certification Examination, a secure computer-based exam that comprised 240 single-best-answer multiple-choice questions that measure residents' cognitive skills (i.e., medical knowledge, diagnostic acumen, and clinical judgment) in general internal medicine. Exam questions provide patient vignettes that require residents to integrate information, prioritize alternatives, and use clinical judgment to reach an appropriate diagnosis or treatment plan. Clinical vignettes have been shown to be a valid method for representing actual clinical practice.[22] The exam was administered in four sections of 60 questions each, with two hours allotted per section. Residents were instructed to answer each question and were not permitted to use external resources to answer questions. There was no penalty for guessing, but questions not answered were scored incorrect. Once residents switched to a new section, they were not permitted to change answers to questions in previous sections. The number of options per question varied but was typically 5. This study used only the 80 questions that explicitly asked the resident to determine the most likely diagnosis.

Participants

On the basis of predictions from a preliminary power analysis, we selected a stratified random sample of 500 of the 9,101 residents who took the exam in 2010 (excluding 44 residents granted special testing accommodations because of disabilities). We acquired data from examinee keystroke log files (i.e., files that record every keystroke) from the testing vendor for a fee. Individual residents were deidentified in the dataset used in the analyses. (Physicians who enroll in an ABIM certification program enter into a business associates agreement that permits the ABIM to use their deidentified data at an aggregate level for research purposes. The ABIM Privacy Policy can be found at www.abim.org/privacy.aspx.) We were blinded to the physicians' identities, and we viewed and analyzed the data in aggregate. Essex Institutional Review Board, Inc., approved this study.

Variables

We developed a computer program to extract residents' response data from the keystroke log files to identify their initial and final responses to each question, and whether the response was incorrect or correct. We also identified the time each resident took to select his or her initial response to each question. For each resident, we computed the number of questions with response changes, and the average seconds spent on initial responses across the 80 diagnosis questions. Reflective diagnostic reasoning was inferred from questions with response changes and longer times on initial responses.

We used residents' overall exam scores to measure their ability (skill level). Overall scores were equated and standardized with a mean of 500 and standard deviation (SD) of 100. We identified residents scoring in the bottom and top quartile of the exam as being in the low- and high-ability groups, respectively. To define case complexity, we first examined the item difficulty values (i.e., the percentage of examinees that correctly answered the question) for each of the 80 diagnosis questions. Higher difficulty values reflect easier questions. On the basis of the distribution of the items' difficulty values, we identified the easiest and most difficult cases by selecting the bottom and top quartile of questions (21 and 20 questions, respectively). Thus, bottom quartile questions had difficulty values < 0.60 and were deemed "difficult," and the top quartile questions had difficulty values ≥ 0.80 and were deemed "easy." We then examined the length, or number of characters constituting these questions. The difficult questions on average contained more characters (mean = 956, SD = 474) than the easy questions (mean = 767, SD = 291). Difficult questions also tended to contain more lab values and clinical test results. Taken together, this suggested that the difficult questions portrayed more complex clinical cases.

Statistical analyses

For each resident, we first identified whether the resident had changed his or her initial response to each diagnosis question. We determined whether the change was from an incorrect to another incorrect response, a correct to an incorrect response, or an incorrect to a correct response. We then tabulated the number and percentage of questions within each of these three categories for each resident. We then computed the mean number and percentage of questions within each category separately for the low- and high-ability residents and for the low- and high-complexity cases. We used contrast *t* tests to assess the differences in the mean number of questions between categories.

We used a hierarchical generalized linear model (HGLM) to assess the relationship between time spent on initial responses to questions and the probability of correctly answering the questions. This model accounts for clustering of exam response data within residents. Item-level characteristics included initial response time, item difficulty, and the time-by-difficulty interaction. Physician-level characteristics included overall exam score (ability), first-time exam taker or repeater status, gender, and medical school graduation country (i.e., U.S./Canada versus international); we included these characteristics in the model if they were statistically significant. Odds ratios (ORs) assessed the odds of correctly answering a question as a function of each variable in the model. In addition, if the time-by-difficulty interaction was significant (i.e., if the impact of time depended on the level of difficulty of the question), we conducted separate regression analyses for the set of low-complexity (easy) and high-complexity (difficult) questions to better understand the nature of the dependency and the role that residents' ability plays in the relationship. In both models, the proportion of items answered correctly was the outcome variable, and residents' average initial response time, ability, and the time-by-ability interaction were included as explanatory variables if statistically significant. Statistical significance was assessed using $P < .05$. Analyses were performed using SAS version 9.3 software (SAS Institute Inc., Cary, North Carolina).

Table 1

Descriptive Statistics for 500 Residents' Initial and Final Item Responses and Frequency of Response Changes on the American Board of Internal Medicine (ABIM) Internal Medicine Certification Examination, From a Study of Diagnostic Accuracy and Reflection, 2010

	Low-complexity (easy) diagnosis questions		High-complexity (difficult) diagnosis questions		All diagnosis questions	
Characteristic	Mean	SD	Mean	SD	Mean	SD
Time on initial responses, seconds	68.6	18.2	97.4	22.6	89.7	25.7
Correct answers on initial response, no. (%)	17.0 (81)	2.5 (12)	9.6 (48)	3.4 (17)	52.0 (65)	16.0 (20)
Correct answers on final response, no. (%)	17.9 (85)	2.5 (12)	10.6 (53)	3.4 (17)	55.2 (69)	15.2 (19)
Questions with responses changed, no. (%)	1.9 (9)	1.7 (8)	3.4 (17)	2.6 (13)	9.6 (12)	5.6 (7)

Results

Of the 500 residents in the sample, 398 (79.6%) took the exam for the first time in 2010; 305 (61%) were male, and 279 (55.8%) graduated from a U.S. or Canadian medical school. The mean overall exam score was 454 (SD = 127). Characteristics of the sample were nearly identical to the 9,101 residents who took the exam in 2010.

Of 500 residents, 498 (99.6%) answered every question and completed each exam section without running out of time. Table 1 presents descriptive statistics for the main variables. As expected, for the high-complexity questions, residents tended to spend more time providing their initial response, changed responses more frequently, and answered more questions incorrectly than for low-complexity questions. On average, residents changed their responses on 12% of all diagnosis questions (or 9.6 questions out of 80).

Figure 1 shows that changing a response from incorrect to correct was approximately twice as likely as changing a response from correct to incorrect. This was statistically significant, except for the low-ability residents changing responses on high-complexity questions.

Table 2 presents the results of the HGLM analysis associating initial response time and other item- and physician-level characteristics with the probability

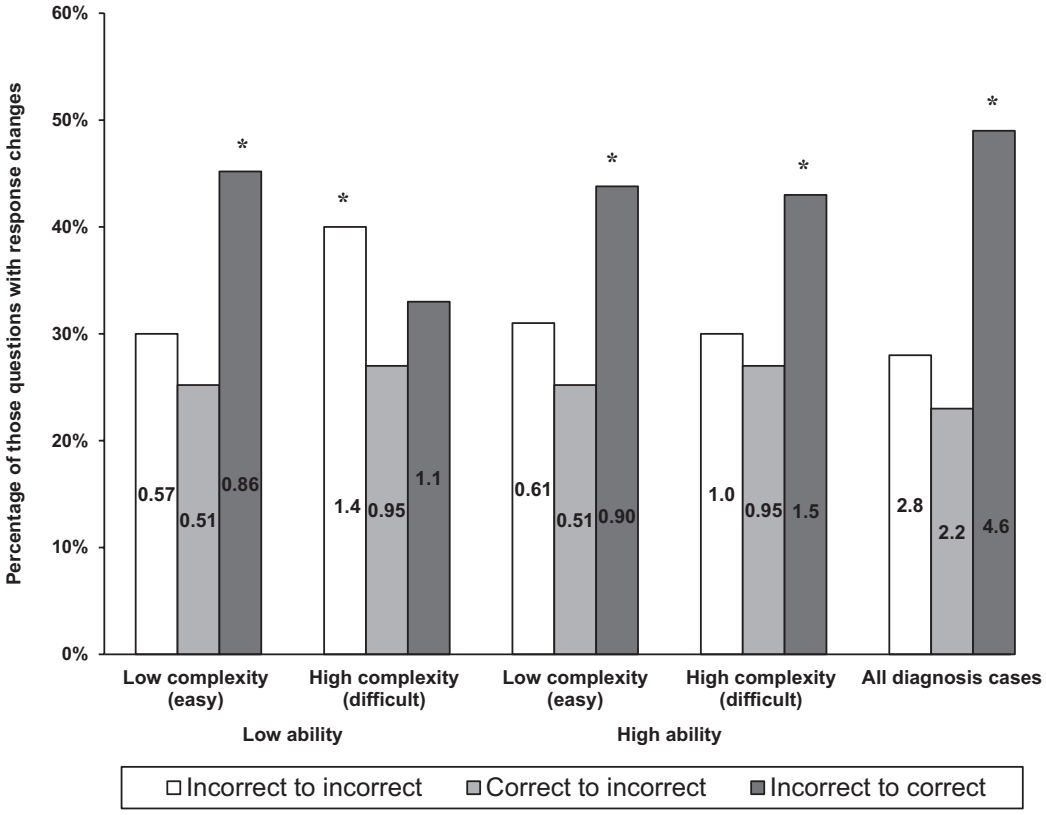

Figure 1 Average number and percentage of questions with each type of response change, by case complexity and ability level, from analysis of 500 residents' responses on the American Board of Internal Medicine (ABIM) Internal Medicine Certification Examination, from a study of diagnostic accuracy and reflection, 2010. Numbers shown within each bar represent the mean number of questions with each type of response change. An asterisk indicates that the mean number is significantly different from the numbers for the other two types of response changes within that category ($P < .05$).

Table 2

Results of the Hierarchical Generalized Linear Model Associating Item-Level and Physician-Level Characteristics With the Probability of Correct Response, From Analysis of 500 Residents' Responses on the American Board of Internal Medicine (ABIM) Internal Medicine Certification Examination, From a Study of Diagnostic Accuracy and Reflection, 2010[a]

Explanatory variables	*b*[b]	Standard error of *b*	*t* value	*P* value	Odds ratio (95% CI)
Fixed effects					
Intercept	0.86	0.02	58.79	< .0001	2.37 (2.30–2.44)
Item-level characteristics					
Initial response time (seconds)	–0.41	0.02	–17.40	< .0001	0.66 (0.63–0.70)
Item difficulty (higher values = easier)	4.74	0.11	45.03	< .0001	114.23 (92.95–140.39)
Initial response time-by-item difficulty interaction	–1.80	0.15	–11.64	< .0001	0.17 (0.12–0.23)
Physician-level characteristics					
Overall exam score (ability)	0.004	< 0.001	35.55	< .0001	1.01 (1.00–1.01)
Medical school country (1 = U.S./Canada, 2 = international)	–0.09	0.03	–3.22	< .001	0.92 (0.87–0.97)
Random effects					
Intercept variance	0.02	0.01	—	—	—
Initial response time variance	–0.02	0.01	—	—	—
Intercept–response time covariance	0.06	0.01	—	—	—

[a]Outcome = probability of a correct response for an individual item (0 = incorrect, 1 = correct).
[b]The regression coefficient, which indicates the change in the probability of a correct response for every one-unit increase in the explanatory variable, controlling for the other variables in the model.

of a correct response on the diagnosis questions. As expected, higher item difficulty values (easier items) were associated with a higher probability of a correct response (OR = 114.23; 95% confidence interval [CI] = 92.95–140.39). Controlling for other variables in the model, greater time spent on initial responses was associated with a *lower* probability of a correct response (OR = 0.66; 95% CI = 0.63–0.70). However, the response time-by-item difficulty interaction was statistically significant (OR = 0.17; 95% CI = 0.12–0.23), indicating that the relationship between time and accuracy was not the same across item difficulty levels. Figure 2 illustrates the nature of the interaction. The inverse relationship between initial response time and accuracy was stronger for easier questions and weaker for more difficult questions. In other words, greater response time was associated with more incorrect answers but to a lesser extent for more difficult cases. Residents with higher overall exam scores (ability) and who had graduated from a U.S. or Canadian medical school tended to achieve greater accuracy on the diagnosis questions.

To further explore the nature of the interaction, we assessed the relationship between initial response time and diagnostic accuracy separately for the low- and high-complexity questions (the extreme cases). Table 3 presents the results of the regression analysis associating average time spent on initial responses with the percentage of correct answers from these initial responses. Consistent with the HGLM analysis, spending *less* time on initial responses for low-complexity questions was significantly associated with more correct answers ($\beta = -0.29$, $P = .003$). However, the significant interaction indicates that this inverse relationship applied only to residents with lower ability ($\beta = 0.41$, $P = .001$). Conversely, for the high-complexity questions, spending *more* time on initial responses was significantly associated with more correct answers ($\beta = 0.14$, $P < .001$), regardless of ability (the interaction was not significant and thus excluded), although the relationship was very modest.

Discussion

We set out to determine whether physicians should trust their initial diagnostic impressions, or whether it may be advantageous to engage in further reflection before making a diagnosis. We found that further reflection, as indicated by residents' changed answers, was beneficial for both simple and complex cases. These results, obtained using realistic clinical scenarios presented in a high-stakes setting, are in agreement with experiments where reflection was encouraged in laboratory-like settings.[13,14] Our approach avoided a problem that has plagued the experimental work on this question to date—namely, determining how to instruct participants to encourage intuitive versus more deliberative responses. In this study, we gave no instructions and simply quantified the residents' natural inclinations.

Our findings based on answer-changing behavior refute the traditional advice given to many test takers to "just trust your intuition." Six other studies over the past century have produced similar results, consistently demonstrating that changing responses is approximately twice as likely to result in a change from incorrect to correct than from correct to incorrect.[23–28] Two factors distinguish our study from these previous ones that enhance its relevance to diagnosis in practice. First, participants were senior medical residents ready to enter practice. Second, considerable progress has been made in producing test questions that are formatted as realistic case vignettes

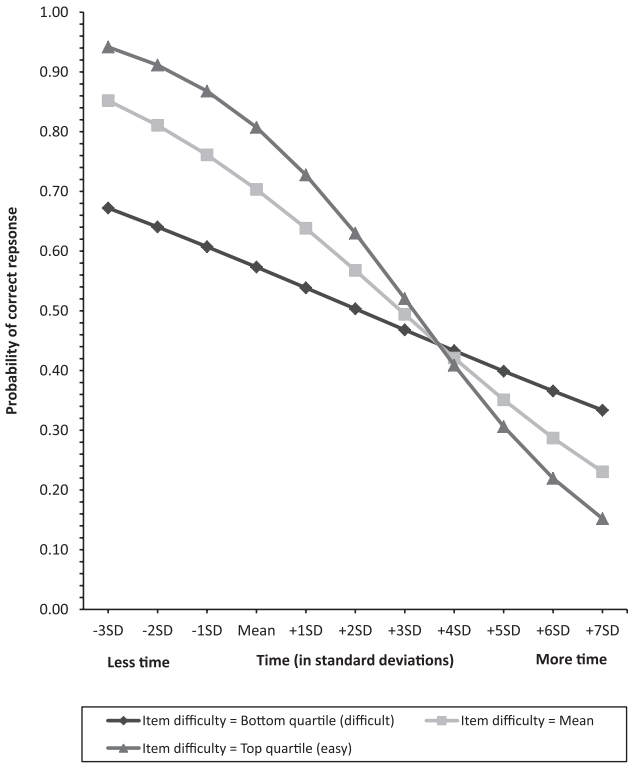

Figure 2 The predicted probability of a correct response by initial response time and item difficulty, from analysis of 500 residents' responses on the American Board of Internal Medicine (ABIM) Internal Medicine Certification Examination, from a study of diagnostic accuracy and reflection, 2010. The line denoted with diamonds represents the relationship for high-complexity (difficult) cases; the line denoted with triangles represents the relationship for low-complexity (easy) cases.

that serve as valid proxies for the study of actual clinical practice.[22,29]

Although our study provides evidence for the benefit of reconsidering the diagnosis, particularly for more complex cases, it remains unclear whether this benefit is derived from conscious thought and reflection, or from the "deliberation without attention" effect described by Dijksterhuis and colleagues.[30] In experimental settings, "deliberation without attention" can be shown to improve intuitive decisions by imposing a time interval between problem presentation and the final decision. Mamede and colleagues[13] reproduced the "deliberation without attention" effect in novice medical trainees presented with simple cases, where introducing a distraction after presenting the case improved the accuracy of their final diagnoses. In contrast, this intervention decreased diagnostic accuracy when novices were given more complicated cases. For more experienced trainees, reflection improved the diagnostic accuracy rate for complicated cases, and for simpler cases the two approaches were equivalent.

Our findings support the idea that reconsidering an original working diagnosis could reduce the likelihood of diagnostic error. Should we recommend this in every case? Strictly speaking, our findings are most relevant to situations where there was an inkling of doubt, as evidenced by the fact that the original response was changed. Our study does not directly address whether a similar benefit would be realized for questions where the clinician was more certain of the original response. Several studies, however, have shown that confidence about a diagnosis is a poor predictor of accuracy,[31–34] and physician overconfidence is one of the important factors underlying diagnostic error.[1] This is because the ease of intuitively retrieving the answer produces a feeling of confidence in the answer, regardless of whether it is actually correct.[35] This feeling of rightness then acts as a cue that further analysis is not needed, leading to high confidence in an erroneous response.[36]

Our study provides additional insight regarding three contextual elements of diagnostic accuracy—namely, the expertise of the clinician, the complexity of the case, and the time needed to establish the provisional diagnosis.

Table 3

The Relationship Between Average Time Spent on 500 Residents' Initial Responses to Questions on the American Board of Internal Medicine (ABIM) Internal Medicine Certification Examination and Percentage of Correct Answers From These Initial Responses, by Case Complexity, From a Study of Diagnostic Accuracy and Reflection, 2010

Explanatory variables	*b*[a]	95% CI for *b*	β[b]	*t* value	*P* value
Low-complexity (easy) diagnosis questions[c]					
Average time on initial responses	–0.002	–0.003 to –0.001	–0.29	–2.94	.003
Ability (overall score)	0.001	0.000 to 0.001	0.39	3.50	.001
Interaction	0.001	0.000 to 0.001	0.41	3.45	.001
High-complexity (difficult) diagnosis questions[c]					
Average time on initial responses	0.001	0.001 to 0.002	0.14	3.79	< .001
Ability (overall score)	0.001	0.001 to 0.001	0.59	16.25	< .001

[a]The unstandardized regression coefficient, which indicates the change in the percentage of correct answers for every one-unit increase in the explanatory variable, controlling for the other variables in the model.
[b]The standardized regression coefficient, which provides a measure of effect size.
[c]The adjusted R^2 for the low-complexity model was 0.55 and for the high-complexity model was 0.35.

Expertise

High-ability residents benefited from further reflection for both simple and complex cases. Low-ability residents did not fare as well on the difficult cases, and although they tended to improve performance by reconsidering their answers even on these questions, the effect was not statistically significant. The results are consistent with the requirement for having a substantial baseline level of competency for reflection to be most effective. Thus, expertise plays an important role in arriving at the best diagnosis initially, and also in refining or revising an initial diagnosis, consistent with previous studies.[37,38]

Time and complexity

The relationship between accuracy and time is complex. We identified three discernible effects of time. First, using all the data, we found that faster response times were associated with a higher likelihood of a correct answer. This makes sense in the context of intuitive responses, which allow for a rapid and typically correct response if the problem is recognized. Second, the data show that high-ability residents answered questions faster than the low-ability group, consistent with their having higher baseline levels of fluency with the exam material. Analogous findings were reported by Sherbino and colleagues[39]; using 25 cases from the Medical Council of Canada Qualifying Exam, greater diagnostic speed was consistently associated with increased accuracy. Both results are consistent with the view that expert clinicians employ rapid and efficient heuristics (hallmarks of intuitive cognition) to solve problems, and they take less time for accurate diagnoses compared with novices.[40] Third, we found interesting effects of case complexity on diagnostic success. For the more difficult questions, the likelihood of a correct answer was proportional to the response time, exactly opposite the effect that was seen for all questions overall. Mamede and colleagues[13] found similar dependence on case complexity; reconsideration led to more accurate diagnoses for more difficult cases, but reduced accuracy for the simpler cases. This finding suggests that a different cognitive approach is used in simple versus complex cases.[41] It is likely that many test takers try to use efficient response strategies and may just quickly guess the answer to difficult questions that they will later return to if they have time. Future research should determine the cognitive factors that help to explain why slowing down to reflect is not particularly advantageous for novices diagnosing easier cases. Moreover, our findings are consistent with the model view that expert clinicians take advantage of intuitive diagnoses when they can, but also know to slow down and increase vigilance as the situation demands.[42,43] In any event, the majority of decisions in clinical medicine are not dependent on short response times. What is relevant to diagnostic ease and accuracy is the degree to which the symptoms of the disease being diagnosed are characteristic ones.[44]

Limitations

Our study has several limitations. First, we do not know to what extent intuition versus analytic thought was used in responding to questions. It is likely that elements of both were involved in the majority of residents' responses. Thus, this study is not simply an evaluation of whether reflection improves on intuition. Many questions answered *without* being changed may also have involved considerable reflection. Our approach also precluded us from knowing if a resident considered changing an answer, but did not.

Second, exam vignettes may have lacked authenticity to practice because clinicians in actual practice may choose to use external clinical resources not available to test takers to make a diagnosis. Third, the spectrum of problems encountered by practicing clinicians may be more or less broad than was reflected in the 80 diagnosis questions. Finally, the residents' behavior could be distorted by virtue of the exam being "high stakes."

Conclusion

The results of this study provide evidence that clinical reasoning can be improved by incorporating further reflection in the diagnostic process, particularly for complex cases. Although our findings apply most specifically to situations where there was some doubt about the initial working diagnosis, we propose that reflection may be beneficial in all cases, given the evidence from other studies that confidence in the diagnosis is not a reliable indicator of its accuracy.

Our study also provides evidence that supports the conceptual validity of the dual-process model of clinical reasoning by demonstrating the divergent effects of time on diagnostic accuracy—that is, excellent accuracy for easier, recognizable problems that can be solved intuitively, but improved accuracy with further analysis and reflection.

Acknowledgments: The authors thank Dr. Jeremy Dugosh for his assistance with retrieval of the exam questions, and Ann Kupinski for writing the computer program to read and obtain the keystroke data from the computer-based exam.

Funding/Support: American Board of Internal Medicine (ABIM) Foundation.

Other disclosures: Dr. Lipner is an employee of the ABIM, as was Dr. Holmboe at the time this study was conducted. Dr. Hess was an employee of the ABIM and is currently a consultant for the ABIM.

Ethical approval: Essex Institutional Review Board, Inc., approved this study.

Previous presentations: Results of this study were presented in abstract form at the Diagnostic Error in Medicine annual meeting, Baltimore, Maryland, November 11–14, 2012.

Dr. Hess is a consultant, Hess Consulting, St-Nicolas, Québec, Canada.

Dr. Lipner is senior vice president of evaluation, research and development, American Board of Internal Medicine, Philadelphia, Pennsylvania.

Dr. Thompson is professor of cognitive psychology, University of Saskatchewan, Saskatoon, Saskatchewan, Canada.

Dr. Holmboe is senior vice president of milestone development and evaluation, Accreditation Council for Graduate Medical Education, Chicago, Illinois.

Dr. Graber is senior research fellow, RTI International, Research Triangle Park, North Carolina, and professor emeritus, at SUNY Stony Brook University School of Medicine, Stony Brook, New York.

References

1 Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. Am J Med. 2008;121(5 suppl):S2–23.
2 Lallemand N. Health policy brief: Reducing waste in health care. Health Aff (Millwood). December 13, 2012. http://www.healthaffairs.org/healthpolicybriefs/brief.php?brief_id=82. Accessed August 6, 2014.
3 Leape L, Berwick D, Bates D. Counting deaths due to medical errors. JAMA. 2002;288:2405.
4 Graber ML, Franklin N, Gordon R. Diagnostic error in internal medicine. Arch Intern Med. 2005;165:1493–1499.
5 Croskerry P. A universal model of diagnostic reasoning. Acad Med. 2009;84:1022–1028.
6 Custers EJ. Medical education and cognitive continuum theory: An alternative perspective on medical problem solving and clinical reasoning. Acad Med. 2013;88:1074–1080.

7 Norman G, Monteiro S, Sherbino J. Is clinical cognition binary or continuous? Acad Med. 2013;88:1058–1060.
8 Gladwell M. Blink: The Power of Thinking Without Thinking. New York, NY: Little Brown and Company; 2005.
9 Kahneman D, Slovic P, Tversky A. Judgment Under Uncertainty: Heuristics and Biases. New York, NY: Cambridge University Press; 1982.
10 Kahneman D. Thinking, Fast and Slow. New York, NY: Farrar, Strauss and Giroux; 2011.
11 Croskerry P. Diagnostic failure: A cognitive and affective approach. In: Advances in Patient Safety: From Research to Implementation. Vol 2. Rockville, Md: Agency for Health Care Research and Quality; 2005:241–254. AHRQ publication no. 050021.
12 Croskerry P. The importance of cognitive errors in diagnosis and strategies to minimize them. Acad Med. 2003;78:775–780.
13 Mamede S, Schmidt HG, Rikers RM, Custers EJ, Splinter TA, van Saase JL. Conscious thought beats deliberation without attention in diagnostic decision-making: At least when you are an expert. Psychol Res. 2010;74:586–592.
14 Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. Med Educ. 2008;42:468–475.
15 Schön D. The Reflective Practitioner: How Professionals Think in Action. New York, NY: Basic Books; 1983.
16 Coderre S, Wright B, McLaughlin K. To think is good: Querying an initial hypothesis reduces diagnostic error in medical students. Acad Med. 2010;85:1125–1129.
17 Moulton CA, Regehr G, Mylopoulos M, MacRae HM. Slowing down when you should: A new model of expert judgment. Acad Med. 2007;82(10 suppl):S109–S116.
18 Trowbridge RL. Twelve tips for teaching avoidance of diagnostic errors. Med Teach. 2008;30:496–500.
19 Ilgen JS, Bowen JL, Yarris LM, Fu R, Lowe RA, Eva K. Adjusting our lens: Can developmental differences in diagnostic reasoning be harnessed to improve health professional and trainee assessment? Acad Emerg Med. 2011;18(suppl 2):S79–S86.
20 Moxley JH, Ericsson KA, Charness N, Krampe RT. The role of intuition and deliberative thinking in experts' superior tactical decision-making. Cognition. 2012;124:72–78.
21 Ilgen JS, Bowen JL, McIntyre LA, et al. Comparing diagnostic performance and the utility of clinical vignette-based assessment under testing conditions designed to encourage either automatic or analytic thought. Acad Med. 2013;88:1545–1551.
22 Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction: A prospective study of 3 methods for measuring quality. JAMA. 2000;283:1715–1722.
23 Bath JA. Answer-changing on objective examinations. J Educ Res. 1967;61:105–107.
24 Davis RE. Changing examination answers: An educational myth? J Med Educ. 1975;50:685–687.
25 Fabrey LJ, Case SM. Further support for changing multiple-choice answers. J Med Educ. 1985;60:488–490.
26 Lowe ML, Crawford CC. First impressions versus second thought in true–false tests. J Educ Psychol. 1929;20:192–195.
27 Mathews CO. Erroneous first impressions on objective tests. J Educ Psychol. 1929;20:280–286.
28 Shahabudin SH. Pattern of answer changes to multiple choice questions in physiology. Med Educ. 1983;17:316–318.
29 Peabody JW, Luck J, Glassman P, et al. Measuring the quality of physician practice by using clinical vignettes: A prospective validation study. Ann Intern Med. 2004;141:771–780.
30 Dijksterhuis A, Bos MW, Nordgren LF, van Baaren RB. On making the right choice: The deliberation-without-attention effect. Science. 2006;311:1005–1007.
31 Friedman CP, Gatti GG, Franz TM, et al. Do physicians know their diagnoses are correct? Implications for decision support and error reduction. J Gen Intern Med. 2005;20:334–339.
32 Podbregar M, Voga G, Krivec B, Skale R, Pareznik R, Gabrscek L. Should we confirm our clinical diagnostic certainty by autopsies? Intensive Care Med. 2001;27:1750–1755.
33 Landefeld CS, Chren MM, Myers A, Geller R, Robbins S, Goldman L. Diagnostic yield of the autopsy in a university hospital and a community hospital. N Engl J Med. 1988;318:1249–1254.
34 Meyer AN, Payne VL, Meeks DW, Rao R, Singh H. Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. JAMA Intern Med. 2013;173:1952–1958.
35 Thompson VA, Morsanyi K. Analytic thinking: Do you feel like it? Mind Society. 2012;11:93–105.
36 Thompson VA, Prowse Turner JA, Pennycook G. Intuition, reason, and metacognition. Cogn Psychol. 2011;63:107–140.
37 Norman GR, Coblentz CL, Brooks LR, Babcook CJ. Expertise in visual diagnosis: A review of the literature. Acad Med. 1992;67(10 suppl):S78–S83.
38 Eva KW, Link CL, Lutfey KE, McKinlay JB. Swapping horses midstream: Factors related to physicians' changing their minds about a diagnosis. Acad Med. 2010;85:1112–1117.
39 Sherbino J, Dore KL, Wood TJ, et al. The relationship between response time and diagnostic accuracy. Acad Med. 2012;87:785–791.
40 Norman G. Research in clinical reasoning: Past history and current trends. Med Educ. 2005;39:418–427.
41 Hodges B, Regehr G, Martin D. Difficulties in recognizing one's own incompetence: Novice physicians who are unskilled and unaware of it. Acad Med. 2001;76(10 suppl):S87–S89.
42 Quirk M. Intuition and Metacognition in Medical Education: Keys to Developing Expertise. New York, NY: Springer; 2006.
43 Rubinstein A. Response time and decision making: An experimental study. Judgm Decis Mak. 2013;8:540–551.
44 Croskerry P, Petrie DA, Reilly JB, Tait G. Deciding about fast and slow decisions. Acad Med. 2014;89:197–200.