Now You See It, Now You Don't: What Thinking Aloud Tells Us About Clinical Reasoning

Judith L. Bowen, MD
Jonathan S. Ilgen, MD, MCR

Assessing diagnostic reasoning is an important, yet messy endeavor. Imagine you are tasked with rating the *reasoning* performance of a senior resident while she conducts a history and physical on a patient with a common complaint. What observable behaviors define expert performance? Can listening to the questions she asks, watching the examination maneuvers she performs, or listening to how she describes her conclusions to the patient reliably capture her reasoning abilities? If a different independent observer was watching the same patient encounter, would the 2 observers reliably reach the same conclusions? Would your ratings of this physician's reasoning skills look the same if she were asked to perform the same tasks on a different patient presenting with vague manifestations of a rare disease?

As these questions illustrate, diagnostic reasoning is not a discrete, enduring, or reliably measurable skill. Accurate measurement requires an observer to interpret processes that are heavily context dependent, rarely articulated, and often occur below conscious awareness of the observed clinician.[1] Not surprisingly, such inferences can be unreliable and prone to substantial rater bias.[2]

In this issue, Heist and colleagues[3] describe qualitative analyses of discourse obtained from first-year residents who were asked to think aloud while solving clinical vignette-based multiple-choice test items. With a small number of novice subjects and just 6 test items, the study provides a relatively narrow slice of the diagnostic process, yet the findings raise some interesting questions. In practical terms, what purpose could think-aloud protocols in this context serve to advance the field of diagnostic reasoning assessment? It depends on who is asking and how the observations will be used.

If the purpose of the think-aloud exercise is to determine test-taking behaviors that lead to correct answers, or what the authors and others have called "test-wiseness,"[4] then knowing and sharing such behaviors could help less "test-wise" residents to achieve higher test scores, something potentially useful for low performers on high-stakes examinations. This strategy is exploited by test preparation businesses designed, for example, to help premedical students improve their MCAT scores.[5] Insights into test-taking behaviors could help level the playing field among test-takers.

If the purpose of the think-aloud exercise is formative assessment, such exercises give residents a chance to explain their reasoning and improve the observer's understanding of how the resident is prioritizing, sorting, and analyzing the information she is given. In this context, the results could be used on a question-by-question basis to correct learners' misinterpretations, identify knowledge or experience gaps, and formulate strategies for additional learning. A variation on this strategy is used in 1-on-1 clinical teaching encounters[6,7] and during hospital ward rounds. Yet, such coaching for learning is still limited to inferences drawn about residents' reasoning processes that are consciously available to them and verbalized.

If the purpose of the think-aloud exercise is to shed light on the relationship between residents' ability to verbalize their reasoning and the accuracy of their diagnoses, previous work analyzing discourse patterns of case presentations may help us understand its value. By analyzing transcripts from oral case presentations, Bordage and Lemieux[8] demonstrated that when clinicians articulated their reasoning for problems where they ultimately arrived at the correct diagnosis, these discourse patterns illustrated language structures that signaled a deep and broad understanding of the clinical problem, and these results correlated with other ratings of diagnostic competence.[9] These authors also demonstrated that, with training and calibration, raters could reliably classify the type of discourse residents used when thinking aloud.[9] Thus, with practice and feedback, faculty supervisors could theoretically be trained to listen for discourse characteristics associated with strong reasoning and diagnostic accuracy as well as for the discourse characteristics associated with weak reasoning and diagnostic failure. They could then intervene with targeted interventions to address learners' specific knowledge deficits in a way that could help them access and apply this learning to subsequent clinical experiences.

Judith L. Bowen, MD, is Professor, Department of Medicine, Oregon Health & Science University; and **Jonathan S. Ilgen, MD, MCR,** is Assistant Professor, Division of Emergency Medicine, University of Washington School of Medicine.

Corresponding author: Judith L. Bowen, MD, Department of Medicine, Oregon Health & Science University, SN-ADM, 3455 SW US Veterans Hospital Road, Portland, OR 97239, bowenj@ohsu.edu

DOI: http://dx.doi.org/10.4300/JGME-D-14-00492.1

If Bordage and Lemieux's[8] discourse classification were applied to the discourse that Heist and colleagues[3] obtained from their think-aloud protocol, would the results be similar? Perhaps. The examples illustrating Heist and colleagues'[3] categories of "reaching closure with difficulty or delay" and "admitting knowledge deficits" sound like Bordage and Lemieux's[8] categories of "dispersed discourse" and "reduced discourse," respectively. Both of these discourse types are associated more often with a failure to reach the correct diagnosis.

Yet we need to be careful about our inferences here: Correlation does not equal causation. Identifying an association between suboptimal discourse and incorrect diagnoses does not necessarily mean that suboptimal discourse represents a faulty diagnostic process that can be remediated. In fact, Nendaz and Bordage[10] also noted that experts were more likely to frame their descriptions as "semantic qualifiers" (peripheral versus central, acute versus chronic), so they studied the impact of teaching second-year medical students how to describe clinical findings in this way. Students in the experimental group were much more likely to use this lexicon, although their diagnostic accuracy remained equivalent to that of peers who used traditional discourse. So "talking like an expert" is not akin to being one: The discourse of experienced clinicians likely signals the extensive knowledge structures they have developed through past rich clinical experiences. Without the backdrop of this knowledge and experience, semantic qualifiers are unlikely to alter novices' diagnostic performance.

Generally, a growing body of evidence challenges the notion that a particular problem-solving strategy will consistently result in diagnostic success,[11–15] and rather than emphasizing a process-based approach to improving diagnostic reasoning, faculty supervisors may respond more effectively to learners who use discourses associated with incorrect diagnoses by addressing underlying knowledge deficits.

If the purpose of the think-aloud exercise is to "improve our understanding of clinical reasoning assessment,"[3] several factors deserve consideration. First, clinicians' diagnostic reasoning can only be inferred.[1] When using think-aloud protocols, the discourse observed reflects complex cognitive processes that are brought to consciousness in the mind of the person thinking aloud. These assessment approaches emphasize "analytic" or System 2 thought processes—the thinking that is consciously available for reflection when residents are prompted to "say everything that goes through your mind as you try to solve each question."[3] However, multiple reasoning processes, both analytic and intuitive (nonanalytic) likely are in play during any reasoning exercise,[16] and it may be the combination of processes above and below conscious awareness that has the highest likelihood of resulting in diagnostic success.[17] Only that which comes to consciousness, however, is available for reflection during problem solving. Inferring that residents use one or the other while thinking aloud is inconsistent with how System 1 processes have been conceptualized[18] and may be prone to substantial hindsight bias.[19]

Second, experience and context matter. How residents from 1 academic health center in the final months of their first year of postgraduate training reason aloud may provide the observer with some clues about their reasoning abilities on a developmental continuum; it could also provide insights about the types of clinical problems encountered in that particular health system. Frequent and recurring encounters with similar patient problems builds strong knowledge structures that enable information to be easily retrieved consciously and unconsciously to solve the next similar clinical problem.[20] If residents are tested on a small number of items that do not reflect their clinical experiences, or that represent problems they have not yet encountered, test performance will suffer and interpreting these test scores as a reliable and valid reflection of residents' reasoning skills would be inappropriate. Some variability in performance can always be attributed to the idiosyncratic experiences of each resident, a problem overcome, for the most part, by broadly sampling a wide range of clinical topics. In practical terms, psychometrically sound multiple-choice tests that can be completed in a reasonable amount of time address this sampling problem.[1] Thus, it is important for clinician educators to be cautious when drawing conclusions about any resident's performance from a small collection of clinical encounters (or in this case, a small number of vignettes).

Third, biases abound, and 2 are worth mentioning. One, an observer's personal knowledge, experience, ability, and personal bias of the importance of a specific element for diagnosis or case management influences his adjudication of a learner's performance in a nonstandard fashion.[2] An experienced supervisor has a way of solving clinical problems unique to his own learning through experience. When a resident's performance does not match the supervisor's idea of a correct way to think (aloud) about a problem, the supervisor may unknowingly judge the resident's performance to be suboptimal. Two, when the observer knows the correct answer to a multiple-choice question a priori—as was the case in the work by Heist et al[3]—he may code the reasoning displayed during a think-aloud protocol as suboptimal when leading to an incorrect answer. Are residents who arrive at the incorrect answer truly suffering from processing errors such as premature closure or overconfidence,[21] or do they simply lack the

knowledge and experience to tackle a particular clinical problem? Residents potentially draw the right or wrong conclusion for many reasons, and if these in-the-moment, think-aloud articulations are truly a teachable process that can improve critical thinking, then adjudication of high/low performance should be divorced from whether the resident's ultimate diagnosis was correct. Otherwise, we should just focus on diagnostic accuracy and ignore how clinicians got there.

References

1 Ilgen JS, Humbert JA, Kuhn G, Hansen ML, Norman GR, Eva KW, et al. Assessing diagnostic reasoning: a consensus statement summarizing theory, practice, and future needs. *Acad Emerg Med*. 2012;19(12):1454–1461.
2 Kogan JR, Hess BJ, Conforti LN, Holmboe ES. What drives faculty ratings of residents' clinical skills? The impact of faculty's own clinical skills. *Acad Med*. 2010;85(suppl 10):25–28.
3 Heist BS, Gonzalo JD, Durning S, Torre D, Elnicki DM. Exploring clinical reasoning strategies and test-taking behaviors during clinical vignette style multiple-choice examinations: a mixed methods study. *J Grad Med Educ*. 2014;6(4):709–714.
4 Rogers WT, Bateson DJ. Verification of a model of test-taking behavior of high school seniors. *J Exp Educ*. 1991;59(4):331–350.
5 Jones RF. The effect of commercial coaching courses on performance on the MCAT. *J Med Educ*. 1986;61(4):273–284.
6 Neher JO, Gordon KC, Meyer B, Stevens N. A five-step microskills model of clinical teaching. *J Am Board Fam Pract*. 1992;5(4):419–424.
7 Wolpaw T, Papp KK, Bordage G. Using SNAPPS to facilitate the expression of clinical reasoning and uncertainties: a randomized comparison group trial. *Acad Med*. 2009;84(4):517–524.
8 Bordage G, Lemieux M. Semantic structures and diagnostic thinking of experts and novices. *Acad Med*. 1991;66(suppl 9):70–72.
9 Bordage G, Connell KJ, Chang RW, Gecht MR, Sinacore JM. Assessing the semantic content of clinical case presentations: studies of reliability and concurrent validity. *Acad Med*. 1997;72(suppl 10):37–39.
10 Nendaz MR, Bordage GE. Promoting diagnostic problem representation. *Med Educ*. 2002;36(8):760–766.
11 Norman GR, Eva KW. Doggie diagnosis, diagnostic success and diagnostic reasoning strategies: an alternative view. *Med Educ*. 2003;37(8):676–677.
12 Mamede S, Schmidt HG, Penaforte JC. Effects of reflective practice on the accuracy of medical diagnoses. *Med Educ*. 2008;42(5):468–475.
13 Ilgen JS, Bowen JL, McIntyre LA, Banh KV, Barnes D, Coates WC, 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(10):1545–1551.
14 Norman G, Sherbino J, Dore K, Wood T, Young M, Gaissmaier W, et al. The etiology of diagnostic errors: a controlled trial of system 1 versus system 2 reasoning. *Acad Med*. 2014;89(2):277–283.
15 Schmidt HG, Mamede S, van den Berge K, van Gog T, van Saase JL, Rikers RM. Exposure to media information about a disease can cause doctors to misdiagnose similar-looking clinical cases. *Acad Med*. 2104;89(2):285–291.
16 Jacoby LL. A process dissociation framework: separating automatic from intentional uses of memory. *J Mem Lang*. 1991;30:513–541.
17 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):405–409.
18 Evans JS. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu Rev Psychol*. 2008;59:255–278.
19 Wears RL, Nemeth CP. Replacing hindsight with insight: toward better understanding of diagnostic failures. *Ann Emerg Med*. 2007;49(2):206–209.
20 Bowen JL. Educational strategies to promote clinical diagnostic reasoning. *N Engl J Med*. 2006;355(21):2217–2225.
21 Berner ES, Graber ML. Overconfidence as a cause of diagnostic error in medicine. *Am J Med*. 2008;121(suppl 5):2–23.