



A THINK-ALOUD APPROACH TO UNDERSTANDING PERFORMANCE ON THE MULTISTATE BAR EXAMINATION

by Sarah M. Bonner, Ph.D.

Editor's note: This article was produced through research conducted by the author that was supported in part by the 2004 Joe E. Covington Award, a research award given annually to a graduate student in testing and measurement by the National Conference of Bar Examiners.

In the summer of 2004, twenty-five law school graduates preparing for the Arizona Bar Examination participated in a study of candidate performance on the Multistate Bar Examination (MBE). The participants were asked to think aloud as they responded to selected test items from a retired test form and their responses were tape-recorded. This technique was used to provide evidence about the substantive cognitive processes that drive examinee response behavior on the MBE. The reports of the participants' thinking were analyzed to provide descriptions of the mental processes used in responding to MBE items. The relationship between test performance and thinking processes was analyzed to determine the relative effect on performance of reasoning from legal principles as compared to other kinds of thinking processes such as general reasoning and use of test-taking strategies.

It was found that the most frequently used cognitive process involved inferences based on the application of legal principles to the fact situations outlined in the questions, and that the tendency to rely on these principles (rather than non-legal reasoning or test-taking strategies) was strongly related to success on the questions. The second most

frequent category involved rehearsal of the facts of the situation.

DEFINING THE "MEASUREMENT CONSTRUCT"

The design of this study was based on the principle that variations in examinees' scores on a given test should correspond to variations in the attribute measured (Borsboom, Mellenbergh, and van Heerden, 2004). It is in part this causal relationship that supports the validity of the proposed interpretation of MBE scores in terms of competence in applying fundamental legal principles to realistic fact situations. The attribute that the test is intended to measure is known as the "measurement construct." Validity addresses the question of whether examinee scores reflect the construct of interest rather than a mixture of factors or some other attribute altogether.¹

Especially in a high-stakes test such as the MBE, it is desirable to minimize the influence on scores of attributes that are considered to be irrelevant to the measurement construct. For instance, the use of test-taking strategies such as guessing and clue-seeking are often considered a source of construct-irrelevant

variance in test scores, and thus an important potential source of invalidity in the measurements (Messick, 1989). Test developers should be aware of any influences that test-taking strategies have on examinees' test scores on the MBE. Users of a test should feel confident that the attributes measured on a test are reasonably related to important skills being tested, especially when important decisions about individuals are attached to the scores, as in the case of the MBE.

For the Arizona study, it was necessary to elaborate on the cognitive processes of interest and to differentiate them from other processes that could be classified as more or less relevant to the construct. According to the 2006 MBE Information Booklet, the MBE is designed to measure the ability to apply to fact situations fundamental legal principles rather than local case or statutory law.²

For the purpose of this study, legal reasoning was defined as the general ability to apply legal rules and principles to specific cases in light of relevant facts. The purpose of the MBE is not to identify examinees with high levels of expertise and deep insights, but to identify, through assessment of certain critical competencies, those who do not possess the basic knowledge needed for entry-level practice. It was decided that general reasoning strategies such as appeals to personal experience, logic, and "common sense" should be distinguished from the construct of legal reasoning.^{3,4}

"The intent in determining critical competencies is to identify areas of knowledge and skill that are critical in the sense that serious deficiencies in a candidate's mastery of these competencies would make it difficult for the candidate to practice law effectively" (Kane, 2005, p. 35). If the test is appropriately

targeted, it would be expected to focus on the skills needed in entry-level practice of law.

THE THINK-ALoud METHOD

In the Arizona study, individual participants were asked to report their thinking orally as they attempted to respond to test items selected from three content areas: Constitutional Law, Contracts, and Torts. This approach is known as the "think-aloud method," or the method of verbal protocol analysis. It has been used in a wide variety of contexts to illuminate cognitive processes underlying behaviors ranging from playing chess to selecting cereals. In the field of educational measurement, the method goes at least as far back as a 1950 study, in which verbal reports of high- and low-scoring examinees on a test of reasoning were compared (Bloom and Broder, 1950).

One might question whether test-takers are able to verbalize in a way that accurately reflects their true mental processes under real test conditions. The legitimacy of this question has been recognized since the foundational years of the study of psychology, when Franz Brentano (1874) stated as a law of psychology the impossibility of accurate inner observation of mental phenomena, on the grounds that the act of observation changes the object of inner perception.⁵

However, considerable evidence supports the use of data gleaned from the analysis of verbal reports.⁶ The current "state of the science" of verbal protocol analysis holds that when verbal reports are taken concurrently with mental processing about cognitive tasks of moderate complexity, useful results can be obtained. In the Arizona study, the think-aloud method was expected to provide relatively complete, accurate, and nonreactive data.⁷

SAMPLE AND ITEM SELECTION

Once the measurement construct had been well-defined and the study method decided, participants were recruited from the Rogers College of Law at the University of Arizona, where candidates for the Arizona Bar Examination were attending BAR/BRI classes and studying in the law library. Eligibility for the study was based on whether prospective participants had completed courses in the relevant subject areas, were planning to take the MBE, and had not already been exposed to the retired test items. The resultant sample was composed of twenty-five law school graduates who were actively engaged in preparing for the July 2004 MBE administration. Slightly more than half of the participants were female, and more than one-third of the participants belonged to an ethnic or racial minority.

Retired test items from a 1998 MBE were obtained from the National Conference of Bar Examiners (NCBE). Rather than investigate response strategies in all six content areas covered on the MBE, the study was confined to three content areas (Contracts, Torts, and Constitutional Law). Because of time constraints (oral reports being a notoriously time-consuming process), only three items in each content area were to be chosen from the pool of 33 to 34 items per subtest. Items that performed particularly well in distinguishing high-achievers from low-achievers⁸ in the 1998 test administration were identified in the selected content areas. From this set of items, a number of those with moderate difficulties were selected to maximize variability in participant responses.⁹

From this small remaining pool, three items in each content area were selected. Then 13 additional items were selected from each of the three content areas for participants to answer silently as a measure

of overall exam performance. The statistical characteristics for the set of nine verbalization items and the set of 39 silent response items from the 1998 test administration were fairly representative of the test as a whole.

THE TEST

The first subset of nine MBE items was administered to each participant individually, after a brief warm-up. Participants were asked to think aloud about their response strategies while being audiotaped, following a standard protocol. Participants were instructed to verbalize all their thoughts without attempting to reflect back upon previous thoughts or to edit their thinking in any way. They were reassured that all thoughts were of interest to the study, and that they should not be concerned about speed of response as they might be when taking the actual MBE, but should try to provide a full description of their mental processes as they attempted to answer the test items without regard to time. It was important to remind participants that speededness of response was not important for the verbalization items, because most participants were well aware of the 1.8-minute average time allotment for MBE items, and were concerned about answering rapidly. But because verbalization is known to increase response time (Norris, 1990), participants' efforts to verbalize within time limits would have necessitated their incomplete analysis of the given problems.

After the administration of the think-aloud item set, participants responded silently to an additional 39 test items, 13 from each of the three content areas. Participants were allowed a limited time of 80 minutes to answer these items, or approximately two minutes per item. This is similar to the amount of time allowed on the actual MBE, which is three hours per 100 items or 1.8 minutes per item. This part of

the study measured the participants' test performances under conditions that simulated actual MBE testing conditions. Following completion of the think-aloud and silent items, participants were asked to categorize the effect of thinking aloud on their ability to respond to the test items as helpful, interfering, or having no effect.

CODING

The audiotaped records of the participants' reported mental processes were transcribed and divided into segments.¹⁰ An example of a section of a segmented protocol is given below, with segment breaks indicated by slashed lines and pauses indicated by ellipses:

"Let me think, would Buyer prevail? . . . //
Buyer didn't go through with the oral condition //
But did they really have to? . . . //
I mean, it was really only in there for their protection, so if they want to skip it, that's . . . //
Can Shareholder use that as grounds to breach?
. . . //
I don't know //
So I'll look at the answers . . . //

The segmented protocols were then coded. In this study, the coding of the segments was based on an inductive, exploratory method described for quantifying verbal data by Chi (1997). This method does not begin with an ideal model and attempt to match results of verbal protocols to the model; rather the analytic process is allowed to proceed with a set of preliminary categories that are used to map the verbal protocols.¹¹ The final system for coding contained the following categories, each of which is

briefly described and then illustrated through two example segments:

1. Rehearsing cues: Segments in which participants paraphrased the fact situation, paraphrased elements of fact provided in the response options, used elements of given facts to justify evaluations without other interpretation or inference, or reinforced memory of facts by underlining or otherwise marking the test form were categorized as belonging to this category.

Example A: *"because that means, oops, no, it's not an oral agreement; it's already in writing . . ."*

Example B: *"even a lifeboat that conformed to statute would not have been launched (laughs) . . ."*

2. Classifying types of problems: Segments in which participants identified or attempted to identify the item's content area (torts, constitutional law, or contracts) were categorized as belonging to this category.

Example A: *"(reads) this looks like a negligence question, or torts, okay . . ."*

Example B: *"there's a fact situation at the top, and then, let's see, some things to assume, it's contracts . . ."*

3. Using inferences based on legal principles: Segments in which participants applied legal principles to elements of the fact situation, referred to legal principles without necessarily applying them to the fact situation, justified evaluations based on legal principles, or interpreted legal principles referred to in the fact situation were classified as belonging to this category.¹²

Example A: *“(reads) uh, that seems to me to be, uh, proximate cause . . .”*

Example B: *“because uh, for—consequential damages are available as far as I know and this seems to be a misstatement of law or rule . . .”*

4. Drawing early conclusions: Segments in which participants made a preliminary hypothesis about the outcome or judgment in a problem prior to reading the response options were classified as belonging to this category.

Example A: *“now as I’m preparing to read the answers I’m thinking that Buyer’s going to win or is going to prevail in his action . . .”*

Example B: *“and I’m going to skip all the way down to unconstitutional ‘cause it just doesn’t seem right . . .”*

5. Making decisions about options: Segments in which participants articulated decisions about choices, such as whether to eliminate or select a response option, were classified as belonging to this category. The category does not include any reasons for the decisions that participants verbalized, which were treated as separate segments and separately categorized. However, statements expressing a preference without a reason were treated as evaluative and classified in this category.

Example A: *“that’s right, that just looks right . . .”*

Example B: *“(reads D) that’s ridiculous; I’m crossing out D . . .”*

6. Using test-taking strategies: Segments in which participants explicitly decided to guess, used deductive elimination strategies, sought clues in item facts and language, or called upon test-taking tactics were classified as belonging to this category.

Example A: *“so I would think this whole part about contributory negligence is just fluff to cloud the issue . . .”*

Example B: *“hm, okay, well, I’m going to try to eliminate some of them . . .”*

7. Making outside inferences: Segments in which participants made inferences that did not involve legal principles and went beyond given facts, often based on common sense or ill-defined intuition, were classified as belonging to this category.

Example A: *“and it doesn’t sound like something the city should be able to zone . . .”*

Example B: *“surely she would have said something to the surgeon . . .”*

8. Non-solution-productive thinking: Segments in which participants verbalized in ways that did not advance their problem-solution were categorized as belonging to this category. Such verbalizations included noncognitive verbalizations, metacognitive verbalizations, and task-irrelevant thoughts.

Example A: *“well, this isn’t a difficult case . . .”*

Example B: *“—just not remembering, I’m trying to run through my outline and I can’t remember wrongful death . . .”*

Each segment that was classified as belonging to the category of rehearsing facts, classifying problems by content area, thinking with reference to legal principles, or drawing early conclusions was also rated as correct or incorrect. The ratings were used to calculate error frequencies for the following types of errors: errors in reading facts, errors in drawing early conclusions, and errors about legal principles.^{13,14}

Statistical comparisons between the proportion of the sample answering the think-aloud (or verbalized) items correctly and the proportion answering the nonverbalized items correctly demonstrated that the verbalized items were more difficult for the participants than the nonverbalized items, but this was also true for the populations that took all the items under standardized testing conditions in 1998.¹⁵ Descriptive statistics on the types of mental processes most frequently found in responses of study participants answering the selected MBE items are provided in Table 1 at right. These statistics are based on dividing the number of segments belonging to each category in each item response for each person by the total number of segments in that person's item response. This operation was performed to place all the thinking processes on a common scale and to minimize the effects of verbosity. To a certain extent, it conveys the efficiency of the participant's thinking processes, or lack thereof, in that a person who applies correct legal principles but has a large number of extraneous or irrelevant

thoughts will have a lower score on "applying legal principles" than a person who applies the same number of legal principles correctly but has fewer extraneous thoughts.

On average, inferences based on legal principles accounted for the largest proportion of all verbalized segments aggregated across items, followed by rehearsing cues from the fact situation and response options. Inferring based on legal principles, rehearsing cues from facts and response options, and making decisions about options were thinking processes considered to be basic to problem solution. However, rehearsing cues from the facts and response options was not considered a highly relevant part of the legal reasoning construct, but instead a manifestation of

Table 1

DESCRIPTIVE STATISTICS ON COGNITIVE PROCESSES AND ERROR TYPES IN VERBAL PROTOCOLS	
Cognitive Processes	Mean Proportion
Inferring based on legal principles	.27
Rehearsing cues	.24
Decision-making	.14
Non-solution-oriented thinking	.18
Test-strategizing	.05
Drawing early conclusions	.03
Inferring beyond givens and principles	.02
Other (not categorized)	.03
Error Types	Mean Frequency
Errors of principles	11
Errors about given facts	1
Errors of prediction	2

Table 2

RELATIVE IMPORTANCE OF DIFFERENT THOUGHT PROCESSES ON PERFORMANCE ON ITEMS	
Independent Variables	Predictive Weight
Legal principles	0.67*
Cues	0.032
Test-taking strategies	-0.13
Early conclusions	0.079
Outside inferences	-0.34*
* Indicates a statistically significant relationship	

refreshing facts in working memory to aid reading comprehension. Also, the somewhat mechanical process of deciding to rule out an option or select it was not of interest in the study, since the inference on which each decision was based was coded separately. These processes represented almost two-thirds of the segments in a person's total set of responses, on average. The standard deviations of these averages indicate that there was considerable variability among participants in the relative proportions of segments in the major categories.

The broad category of non-solution-oriented thinking comprised 18 percent of the segments for participants on average. The nature and preponderance of these segments (a typical segment classified as non-solution-oriented would be "I'm trying to think," or "I just can't remember . . .") suggested that participants were doing considerable mental processing that they were not able to verbalize concurrently.

To assess the relative effect on overall performance of the proportions of different types of cognitive processes, three analyses were conducted. Results

were similar across these analyses. The outcome measure of overall performance in these analyses is a composite score, based on both the verbalized and nonverbalized item scores. The construct-relevant predictor was the proportional use of inferences based on legal principles; the construct-irrelevant thinking processes were proportional rehearsal of cues (related to reading comprehension), making inferences outside facts and principles, and using test-taking strategies. Drawing early conclusions was an unpredicted strategy, and no hypothesis was made about the effect of this thinking process. The model was statistically significant and accounted for a large proportion of variance in the dependent variable. Using inferences based on legal principles was a statistically significant predictor of performance, with a positive coefficient. The use of inferences that went beyond the facts and principles was also a significant predictor in some analyses, but in this case with a negative coefficient. Results are summarized in Table 2.

Analyses of the contribution of different types of errors to performance supports the importance of correct construct-relevant knowledge and the irrelevance of variability in reading comprehension.¹⁶ These analyses also showed the strong negative impact of drawing early erroneous conclusions. The results of this analysis of error types are provided in Table 3. The model accounted for a large proportion of variance in scores. As predicted, the category of errors of principle was a statistically significant predictor, while errors of fact accounted for little variability. The category of errors in drawing early conclusions was also found to be a statistically significant negative predictor.


Table 3

SUMMARY RESULTS OF REGRESSIONS OF ERROR TYPES ON MBE ITEM PERFORMANCE	
Types of Errors	Predictive Weight
Errors in legal principles	-0.73*
Errors about given facts	0.19
Errors in early conclusions	-0.55*
* Indicates a statistically significant relationship	

CONCLUSION

The results of this study support the validity of the selected MBE item scores as measures of that part of the construct of legal reasoning defined as the application of legal rules and principles to specific cases. Different analyses converged on strikingly similar results: when proportional use of different types of cognitive processes was used to form predictors from the verbal reports of participants' thinking, the application of legal principles had by far the strongest positive relationship to performance. High performance on the selected items was associated with greater proportional use of construct-relevant thinking processes (making inferences based on legal principles) and had no significant relationship to potential sources of construct-irrelevant variance, particularly test-taking strategies.¹⁷

The study reported here is limited in that it involved a relatively small sample size and used items from only three of the six content areas on the MBE. Nevertheless, the results indicate that most of the variability in item performance is accounted for by variability in the effective use of legal principles in analyzing fact situations, and that potentially irrelevant constraints like reading comprehension,

test-taking strategies, and item cues have relatively little impact. 

REFERENCES

Bloom, B. S. and L. J. Broder (1950). *PROBLEM-SOLVING PROCESSES OF COLLEGE STUDENTS*. Chicago: University of Chicago Press.

Borsboom, D., G. J. Mellenbergh, and J. van Heerden (2004). *The Concept of Validity*. *PSYCHOLOGICAL REVIEW*, 111(4), 1060–1071.

Brentano, F. (1874). *PSYCHOLOGY FROM AN EMPIRICAL STANDPOINT* (trans. A. Rancurello, D. B. Terrell, and L. L. McAlister, 1973). New York: Humanities Press.

Chi, M. T. H. (1997). *Quantifying Qualitative Analyses of Verbal Data: A Practical Guide*. *JOURNAL OF THE LEARNING SCIENCES* 6(3), 271–315.

Ericsson, K. A. and H. A. Simon (1993). *PROTOCOL ANALYSIS: VERBAL REPORTS AS DATA* (rev. ed.). Cambridge, Mass.: MIT Press.

Kane, M. T. (2005). *The Role of Licensure Tests*. *BAR EXAMINER* 74(1):27–38.

Messick, S. (1989). *Validity*. In R. L. Linn (Ed.), *EDUCATIONAL MEASUREMENT* (3d ed., pp. 13–103). Washington, D.C.: American Council on Education and National Council on Measurement in Education.

Norris, S. P. (1990). *Effect of Eliciting Verbal Reports on Thinking on Critical Thinking Test Performance*. *JOURNAL OF EDUCATIONAL MEASUREMENT* 27(1), 41–58.

ENDNOTES

1. For example, in the content area of mathematics, a test composed of word problems might be developed to measure math problem-solving skills. If a validation study revealed a strong relationship between scores on the test and vocabulary, the validity of the interpretation of the test scores in terms of math problem-solving only would be called into question. Such results might occur even if each problem on the test clearly related to mathematical content.
2. While the MBE covers topics in six areas of the law—contracts, torts, constitutional law, evidence, criminal law, and real property—empirical data show that examinees proficient in one content area tend to be proficient in all content areas. A review of research on the MBE found that “MBE subtest scores are so highly related to each other that differences between a candidate’s scores on two subtests are more likely to be due to chance than to systematic differences in the candidate’s ability in these areas” Klein, S. P. (1993). *SUMMARY OF RESEARCH ON THE MULTISTATE BAR EXAMINATION*, p. 18.

3. Previous studies have not provided strong empirical evidence that MBE scores measure something distinct from general reasoning. For instance, Klein (1993) cites a 1992 study that found that in thirteen of the fifteen largest law schools in California, performance on the MBE correlated somewhat more strongly with comparatively direct measures of legal skills than with performance on the LSAT, a measure of more general reasoning. Klein contends that the difference in these correlations is evidence of the validity of the MBE as a test of “developed legal ability” as opposed to skill in taking multiple-choice tests. However, the differences in the reported correlations are not large, and the comparative weakness of the MBE-LSAT correlation may be a function of the span of time between the two measures. Klein also reports a study comparing the performance of first-year California law school students with the performance of law school graduates. The fact that the graduates did better than novices is inconclusive evidence that general reasoning skill is not the construct underlying MBE scores. Plausible alternative hypotheses such as motivation and maturity may explain some of the differences in performance between novices and law school graduates.

4. Other than analyses of MBE scores like those cited above that have attempted to distinguish legal reasoning from general reasoning, little empirical research has been done on the kinds of reasoning strategies and mental processes that are relevant to the practice of law. Discussions about the thinking processes of lawyers are mostly historical, descriptive, or prescriptive.

A review of various law school texts revealed that today’s mainstream legal reasoning combines the relatively mechanical application of rules to straightforward cases, with a modern recognition that rules exist in historical contexts and can change according to policy needs. Modern lawyers appear to use a mixture of dogmatism and pragmatism on a case-by-case basis. However, traditional values are still upheld, especially for pedagogical purposes. The tendency to maintain at least a partially traditional and dogmatic approach is further evidenced by law teachers’ fairly widespread use of a method known as IRAC (Issue-Rule-Application-Conclusion) as a heuristic for teaching the structure of legal analysis.

5. The principal modern arguments against the use of people’s self-reports of mental processes are given by Nisbett and Wilson (Nisbett, R. E. and T. D. Wilson. (1977). *Telling More Than We Can Know: Verbal Reports on Mental Processes*. PSYCHOLOGICAL REVIEW 84(3), 231-259). They state “the accuracy of subjective reports is so poor as to suggest that any introspective access that may exist is not sufficient to produce generally correct or reliable reports” (p. 233). They cite extensive experimental research in psychology in which research participants failed to report, inaccurately reported, or even denied recognition of mental processes that had been induced in them by the experimenter. For example, in an experiment demonstrating their argument, Nisbett and Wilson asked passersby to choose the best quality pair of nylon stockings from a selection of four identical pairs. The passersby demonstrated a position effect in their choices: they were almost four

times more likely to choose the stockings on the far right than to choose those on the far left. When asked to report the reason behind their choices verbally, the passersby never mentioned the stockings’ position. In fact, when they were asked if position might have had an effect on their choice of stockings, they all denied the possibility.

6. Ericsson and Simon (1993) review multiple experimental and quasi-experimental studies showing no effect on performance for verbal reports of mental processes when subjects verbalize concurrently with task performance without introspection. In fact, according to a recent review of the arguments about verbal reports (Pressley, M. and P. Afflerbach. (1995). *VERBAL PROTOCOLS OF READING: THE NATURE OF CONSTRUCTIVELY RESPONSIVE READING*. Hillsdale, N.J.: Lawrence Erlbaum.), one of the main contributions of Ericsson and Simon’s (1993) argument is that their perspective “is consistent with even apparent antagonists of verbal reports” (p. 8). Because Ericsson and Simon stress the verbal report of information held in short-term memory only, which can be derived directly by external stimulation or by cueing associations held in long-term memory, their perspective provides an explanation of why certain kinds of verbal reports are likely to be inaccurate.
7. This was in part because the specific participant instructions and procedure used met the requirements for a concurrent verbalization as described by Ericsson and Simon (1993). Also, the cognitive tasks did not involve responding to reportedly problematic stimuli such as simple recognition, or pictorial or spatial thinking. Furthermore, the problems attempted by participants in the Arizona study involve a specific culture with clearly prescribed rules for input and output. Even Nisbett and Wilson (1977), who generally oppose the use of verbal reports, concede that in such cases verbal reports may be accurate and useful.
8. This is known as the “point-biserial correlation,” a measure of item discrimination, or the “ability” of the item to distinguish high-achievers from low-achievers.
9. The difficulty of the item is the measure of the proportion of respondents who are able to answer it correctly, in this case between 35 and 65 percent.
10. A detailed discussion with examples of different approaches to the segmentation process is provided by Chi (1997). For the present study, a fine grain size was chosen in order to maximize the amount of information obtained from the protocols about the presence or absence of various kinds of thinking. Segments were defined as meaningful utterances separated by syntax of grammatical subordination or long pauses, except where participants were paraphrasing givens from an item’s fact situation or options. In those cases, the entire paraphrase was treated as one unit even if it included multiple subordinated ideas, unless long pauses separated parts of the paraphrase.
11. Preliminary categories developed for coding the protocols included the following: gathering facts, thinking with reference to specific cases, thinking with reference to legal princi-

- ples, hypothesis development, evaluating options, guessing, and noncognitive verbalizations. When the preliminary categories were found not to cover the protocols adequately, additional codes were created or the preliminary codes modified, until a coding system was developed that covered the protocols adequately. Chi calls this process “piloting the analyses” (p. 8). Once a reasonable system for coding was developed, the rest of the segments were coded.
12. The sub-category of thinking with reference to specific cases was considered a part of using legal principles, due to its very low incidence (three occurrences over approximately 4,000 coded segments). With that addition, legal principles were those listed by the subject matter outlines provided by NCBE or named by subject area experts from the Rogers College of Law at the University of Arizona who were consulted to explicate the problem solutions.
 13. These types of errors are very similar to those analyzed by Elstein, Shulman, and Sprafka in their studies of medical problem-solving, which included misinterpretation of data cues, inaccurate hypothesis generation, and errors in evaluation of hypotheses. Elstein, A. S., L. S. Shulman, and S. A. Sprafka. (1978). *MEDICAL PROBLEM SOLVING: AN ANALYSIS OF CLINICAL REASONING*. Cambridge, Mass.: Harvard University Press.
 14. Twenty percent of the transcripts were coded by a second rater to assess interrater agreement on coding. Transcripts were systematically selected for the interrater agreement study to ensure that the second rater rated every item and every participant. Transcripts were randomly ordered within items, and participant identification was removed to reduce possible sources of bias in the ratings. Initial interrater agreement was about 70 percent averaged across items and persons, so that the rating system appeared fairly unreliable. Examination of the areas of disagreement revealed that the category for early concluding was not being used by the second rater due to lack of clarity in training. More importantly, the second rater had a more limited range of terms to which the category relating to legal principles applied, and was therefore underusing that category. For instance, a segment in which the term “cause” was inferred was categorized by the second rater as an inference outside legal principles, because “cause” was taken in its common meaning rather than as a legal term. Once the problems that arose due to differences between the raters’ codings were diagnosed, the raters met for further training. Using the NCBE content outlines and expert problem-solving models developed with the assistance of law school faculty, a comprehensive list of legal principles potentially relevant to the nine items was devised. Then the segments were rescored by each rater. Results of the interrater agreement analysis averaged across all items and all processes were considered to be acceptable and indicative that the mental process categories of interest could be used consistently and non-idiosyncratically to categorize segments of participant responses.
 15. The known population difficulty based on the 1998 test administration was found to predict the item difficulties obtained in the present study fairly well, and there was no indication that verbalization made the items appreciably more difficult.
 16. Three types of errors had been coded for use as predictors: errors about legal principles, errors about facts in the item situations, and errors in drawing early conclusions. The error variables were derived from frequency counts of the different types of errors as classified based on the verbal protocols. Frequencies rather than proportions were used because errors were seldom exactly repeated, and two errors within one response, even if of the same type, were considered to be substantively different from a single error. Of these three types of errors, errors about legal principles were predicted to impact performance. Errors about given facts were not expected to relate to test performance, because if the test primarily measures legal reasoning, basic reading comprehension should not be an important source of variability in scores. Because drawing early conclusions was an unexpected thinking process found during the coding of the segmented verbal protocols, no prediction was made about the effect of mistakes in doing so.
 17. The making of inferences outside facts and principles represents a nonlegalistic tendency to reason from beliefs rather than evidence. This tendency, too, was found to be associated with low performance on the selected items.



SARAH M. BONNER is an assistant professor in the Department of Educational Foundations and Counseling Programs at Hunter College, CUNY. She received her Ph.D. from the Department of Educational Psychology at the University of Arizona in 2005. This article is based on her research that won the 2004 Joe E. Covington Award. The author would like to acknowledge the support of the National Conference of Bar Examiners.