AN INVESTIGATION OF THE SUBSTANTIVE PROCESS VALIDITY OF MULTISTATE BAR EXAMINATION ITEMS THROUGH VERBAL PROTOCOL ANALYSIS

by

Sarah M. Bonner

_____

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2005

THE UNIVERSITY OF ARIZONA

GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation

prepared by Sarah M. Bonner

entitled An Investigation of the Substantive Process Validity of Multistate Bar Examination Items through Verbal Protocol Analysis

and recommend that it be accepted as fulfilling the dissertation requirement for the

Degree of Doctor of Philosophy

_____ Date: May 17, 2005
(Jerome V. D'Agostino)

_____ Date: May 17, 2005
(Darrell L. Sabers)

_____ Date: May 17, 2005
(Anthony J. Nitko)

_____ Date: May 17, 2005
(Lee Sechrest)

_____ Date: May 17, 2005
(Patrick McKnight)

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: (May 17, 2005)
Dissertation Director: (Jerome V. D'Agostino)

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

TYPE NAME HERE: Sarah M. Bonner

ACKNOWLEDGEMENTS

Without the influence of my husband, Rahul, I never would have undertaken this endeavor. Also my son, Aidan, has encouraged me with the example of his optimism, and has patiently borne my complaints.

I gratefully acknowledge the support of my committee members, Jerry D'Agostino, Darrell Sabers, Tony Nitko, Lee Sechrest, and Pat McKnight. Each of them has contributed in important ways to increasing my understanding of psychological measurement. Though I still have abundant things to learn as an educator and researcher, they have set me on a path that I look forward to with excitement. It has been a pleasure to learn from people with so much expertise and enthusiasm. In particular, I would like to mention that Darrell Sabers taught me the wisdom of working in pencil. Thanks too, to Keith Meredith, who has an excellent shoulder for crying on.

I want to express my thanks to the students, staff, and faculty at the Rogers College of Law, especially those faculty members who helped me understand an unfamiliar field of expertise. And I would like to thank Michael Kane and the National Conference of Bar Examiners, whose funding and encouragement initiated this study.

Finally, my parents—my mother, who has always been a model of clarity, reason, and correctness, and my father, who was a wonderful teacher, a good person, and a skeptical inquirer.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

The dissertation describes a think-aloud study that investigated the internal processes of 25 participants responding to selected items from the Multistate Bar Exam (MBE). The MBE is a nationally administered high-stakes licensure test used in 53 U.S. jurisdictions. The following questions were addressed by the study: Which mental processes are most frequently used in answering the selected items? What is the relative effect on performance of reasoning from legal principles as compared to general reasoning or use of testwiseness principles? When variance in performance due to similarity of responses to expert problem-solving models is accounted for, do divergent mental processes add to prediction? Do minority ethnic examinees use different cognitive processes or commit different types of errors when answering items in selected MBE content areas as compared to majority ethnic examinees?

The dissertation discusses the construct of legal reasoning, the method of verbal protocol analysis, and potential method effects of reactivity and veridicality. The study method, including the development and reliability of the rating systems used to quantify the verbal transcripts for analysis are also described.

Findings of the study are presented. Using legal principles and rehearsing facts were the most commonly used mental processes. Participants used a number of other mental processes which are described. Using legal principles and avoiding extraneous inferences were associated with increased performance. Drawing premature conclusions and making errors in legal principles negatively affected performance. The effects of different error types and construct-

irrelevant thinking such as cue-using strategies are also be reported. Similarity of responses to expert problem-solving models predicted performance on verbalized items, and adding divergent mental processes as predictors did not significantly increase the amount of variance accounted for. No evidence was found in the verbal responses that minority ethnic examinees used different cognitive processes when answering items in selected MBE content areas as compared to majority ethnic examinees; however, a method effect appeared for the minority examinees in the study that complicates interpretation. Some evidence was found of differences in response patterns for minority examinees that could not be explained by the verbal responses.

CHAPTER 1

INTRODUCTION

The *Multistate Bar Exam* (MBE) is a nationally developed and administered test used for high-stakes certification decisions about individuals. It was developed by the National Conference of Bar Examiners (NCBE) and has been used since 1972 to measure broad abilities to apply fundamental legal principles, abilities which include analyzing legal relationships arising from a fact situation, taking positions as an advocate, and making judgments about interpreting, drafting, or counseling. Items on the MBE assess these abilities in six content areas: contract law, real property, rules of evidence, criminal law, torts, and constitutional law.

Taking the MBE is part of the process for "passing the bar" in most United States' jurisdictions. "Passing the bar" is the process by which candidates are formally admitted to the practice of law. In 2003 the MBE was used in 53 U.S. jurisdictions (all jurisdictions except the state of Washington, Louisiana, and Puerto Rico). The MBE is usually administered in combination with an essay examination, the *Multistate Professional Responsibilities Exam* (MPRE), and, in some jurisdictions, a performance assessment task. Individual jurisdictions set their own passing scores on each separate component of the bar exam or on combined scores, and assign their own weight to each component. In most jurisdictions the various assessments are placed on a common scale and weighted, with the MBE score weight ranging from one-third to one-half. Thus low scores on the MBE can be partially offset by high scores on

the essay examination, for instance. Among the ten jurisdictions that required a minimum passing score on each separate exam in 2003 (when the national scale score mean on the MBE was 140), all but the Northern Mariana Islands and Palau required a scale score between 130 and 135. The MBE is considered a very high stakes test by applicants to the bar, who commonly spend several thousands of dollars in bar exam preparation courses.

The present study uses analysis of the cognitive processes that underlie examinee performance as evidence about the validity of the MBE as a measure of the ability to apply legal principles. Thus it represents an instance of the substantive process approach to validation of test score interpretations in terms of constructs. Verbal reports of examinees' thoughts as they respond to multiple-choice items from selected content areas of the 1998 form of the *Multistate Bar Exam* (MBE) are used to access the cognitive processes that underlie examinee performance. This method has traditionally been referred to as the "think-aloud" method, or the method of verbal protocol analysis. Ericsson (1987) provides the rationale for this method when he discusses the likelihood that the same or similar thinking processes are used by examinees to solve test problems of specific types, and the usefulness of verbal reports for accessing such processes. He holds that "it is reasonably likely that situations will be found where equivalence of the cognitive processes on the different test items is a good approximation. Through the collection of verbal reports on the cognitive processes used on test items, it is possible to identify blatant violations of the assumptions of measurements of the same

general process sequence, by identifying systematically different strategies among the tested subjects" (Ericsson, 1987, p. 203).

In the case of the MBE, investigation of examinee response processes provides information about the contributions to test performance of thinking relevant to context-specific legal principles as compared to general reasoning. While evidence mounts in favor of the predominant value of context-specific content knowledge for success in fields such as physics and medicine (Perkins & Salomon, 1989), it has yet to be demonstrated whether "thinking like a lawyer" means using principles and content specific to the legal domain, or more general reasoning skills. Although the MBE does not claim to be a comprehensive test of legal reasoning, information from analysis of processes used on the test provides evidence about the types of reasoning most related to performance in solving certain limited types of legal problems that are considered prerequisite to entry into practice.

Insight into response processes can help users understand if any areas on the MBE are particularly sensitive to construct-irrelevant response strategies. The use of testwiseness principles such as guessing and clue-using strategies are considered a source of construct-irrelevant variance in test scores, and thus an important potential source of invalidity with respect to construct interpretation of the observed scores (Messick, 1989). Test developers should be aware if test-wiseness strategies have a significant influence 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 constructs in the domain, especially when important decisions about individuals are attached to the score, as in the case of the MBE.

Moreover, possible bias in testing as a source of invalidity is of increased concern when stakes are high, as in licensure testing, because high-stakes decisions based on biased test interpretations can result in inequities in professional fields such as law. Klein (1993) reports that examinees belonging to racial or ethnic minorities (Asians, Hispanics, Blacks) had lower scores on the average on the MBE compared to White examinees in a 1992 study of California examinees. Significant differences have also been found between white and nonwhite examinees in first-time pass or fail overall bar passage outcomes as well as eventual pass rates (Wightman, 1998). For first-time test takers, the performance differences are considered to be of both practical and statistical significance. While failures among black first-time test takers make the largest contribution to the analysis of performance differences, "differences between pass rates for whites and pass rates for other ethnic groups are substantial" (Wightman, 1998, p. 28) and remain statistically significant when blacks are omitted from the analysis.

This is not to imply that the MBE is necessarily biased. The differences were found by both Klein and Wightman to be highly correlated with differences in academic proficiency among groups as measured by the LGPA and LSAT scores. When such academic achievement and ability measures were used as statistical control variables by Klein, information about examinees' race or ethnicity did not significantly improve the prediction of MBE scores. This supports the conclusion that MBE test items are not biased and do not exacerbate differences associated with race or ethnic group. It does not, however, provide insight into the behaviors or cognitive differences that cause

the observed differences in performance among different groups of examinees. Examining the types of errors made by different subgroups may help diagnose the causes underlying such gaps. One of the goals of this investigation is to provide insight into the behaviors or cognitive differences that cause the observed differences in performance among different groups of examinees.

As an example of this approach to examining internal processes in population subgroups to explain observed score differences, Bond (1990) used concepts from information-processing theory to attempt to explain poor performance by black students taking the quantitative reasoning section of the SAT. His work revealed no differences in types of errors or misconceptions between majority and minority ethnic problem solvers on SAT Math items.

Finally, the study adds to the body of evidence about the use of verbal protocols for test validation. Because of the growing emphasis on cognitive approaches in assessment (Norris, 1990), more studies of the usefulness of the analysis of verbal protocols for validating test scores are needed. Multiple methods for performing quantitative analysis on the rich data derived such from verbal protocols is demonstrated, using both exploratory and confirmatory techniques. Even though verbal protocol analysis is increasingly advocated as a means to evaluate the use of educational measures (Leighton, 2005), few studies have described in detail the application of the method and usefulness of the results. Therefore, it is hoped that this aspect of the study makes a methodological contribution to the field.

Background

From the time of the American Revolution until the mid-20th century in the United States, passing the bar was rarely based on formal examination. Prior to the Revolution, each colony set its own standards for admission to the bar, usually based on a long period of apprenticeship and formal exams (Friedman, 1985). The emerging colonial legal system was strongly influenced by English tradition, although a graded system along the lines of the barrister-solicitor model did not survive the Revolutionary period in the colonies where it was tried. While standards for admission to practice may have been rigorous in some colonies, they were not consistent. In Massachusetts, for instance, five years' apprenticeship was required and each court had the authority to grant admission to its own lawyers. In Virginia, the high court had control over licensing and admission to the bar. In Rhode Island, any court could grant admission, and candidates, once admitted, could practice in any other court in the state. Throughout the colonies, practicing lawyers were few in number, and had no control over bar admissions.

At the time of the American Revolution, many leading lawyers who were Loyalists left the country. The number of young men seeking entry into the legal profession increased substantially, as did demand for their services. Bar admissions began to be decided more idiosyncratically by individual lower courts. This led to increasing laxity in admissions processes in the early 19th century, when "the lawyer who could get himself admitted, even by the most slipshod local court, was a fully licensed member of the bar of the state, and could practice before any court" (Friedman, 1985, p. 316). Admission to the

bar in the mid-1800s was usually based on oral examination, and anecdotal evidence indicates that particularly for those lawyers willing to practice in western territories, examination might consist of a few informal questions while walking with a judge in the countryside or sitting in a local tavern. In the 1840s a few states eliminated all requirements for admission to the bar except good moral character.

The late 1800s saw the rise of law schools as an alternative to legal apprenticeship. Friedman (p. 607) cites an increase from 15 American law schools in operation in 1850 to 102 in 1900. Matriculation from a law school, however, did not necessarily lead to higher standards for admission to the bar. Many states granted their state public universities a "diploma privilege," meaning that any graduate from the university's school of law was automatically admitted to the state bar without examination. Because both within and between states law schools varied in the quality of legal preparation they offered, the diploma privilege did little to increase professional standards.

Perhaps in part as an attempt to control the flood of applicants coming out of law schools, between 1870 and 1890 there was a drive to tighten standards for admission to the bar. With the support of the newly-founded American Bar Association (ABA) which opposed the diploma privilege, written examinations became increasingly common. Before the ABA was founded in 1878, only Massachusetts and New York used written bar exams. "By 1917, centralized boards of bar examiners existed in thirty-seven jurisdictions" (Stevens, 1983, p.99). In 1931 the National Conference of Bar Examiners (NCBE) was founded under the sponsorship of the ABA. The NCBE's initial objectives "indicated an . . . orientation,

namely toward standardizing legal education and admissions to the bar nationally" (Stevens, p. 177). The two goals of developing uniform standards for education in the practice of law, and standardizing the bar examination process, continue to be central objectives for the NCBE. The widespread use of the MBE in the United States indicates the progress that has been made towards accomplishing the latter objective.

The purpose of standardized educational and testing requirements for the legal profession historically was and continues to be "to provide protection against the risks inherent in admitting candidates who lack the basic knowledge, skills, and judgment necessary for safe and effective practice. They are designed to ensure that the candidate has achieved a reasonable level of competence in applying professional knowledge, skills, and judgment to practice problems. Without such competence, new practitioners would likely make mistakes that could put their clients at risk" (Kane, 2005, p. 27). With the MBE, the NCBE has taken the approach of using a standardized multiple-choice examination to measure basic critical competencies. Two important inferences are to be made about the test scores: one, that differences in scores are valid indications of different levels of competence in the construct of applying legal knowledge, skills, and judgment; two, that the standard set for passing classifies test-takers appropriately according to whether they are minimally competent. The validity of the first of these inferences is to be investigated in this study.

It is important to validate test score interpretations of important tests such as the MBE in terms of the measurement construct, because a basic principle of testing and measurement is that variations in examinees' scores on a given test should be caused by corresponding

variations in levels of the attribute measured (Borsboom, Mellenbergh & van Heerden, 2004). It is in part this causal relationship that allows users to consider the use of a given test valid, along with evidence that the use of the test does not result in unfair or adverse unintended consequences. The difficulty lies in demonstrating the causal relationship. Since researchers can not randomly assign examinees to levels of an attribute, the presence of the relationship may be inferred from sources of evidence such as non-experimental observational or correlational studies, logical analysis, and quasi-experimentation.

There are some special cases of test validation where researchers may place a reduced emphasis on making causal inferences. When the proposed test use is prediction of a criterion, the use of the test scores may be considered valid if they correlate highly with the criterion, without inquiry into the causal relationship. However, in the case of a professional licensure examination such as the MBE, lack of a clear external criterion makes this type of validation difficult. Those who pass the bar go on to specialize, so broad abilities to apply fundamental legal principles become irrelevant to legal practice once the bar is passed. Few if any practicing lawyers in U.S. jurisdictions ever need to use principles across all six tested content areas. Subscores in a content area might be validated with reference to a criterion measure administered to specialists in that content area, although such criterion measures of excellence in different areas of the law do not currently exist. Even if they did, as with any test that serves as a gateway to practice, the outside criterion would measure

only a restricted range of the total set of scores to be validated; people failing the bar would not be measured on the criterion.

The causal link between attribute and test scores may also be considered less important than the issue of how well the test items sample the achievement domain, when the test scores are to be used as descriptors of absolute or relative achievement as in many educational testing situations. In such situations, panels of experts use logical judgmental methods to assess individual items for content relevance, and the test as a whole for the degree to which it representatively samples the given curriculum or domain. The NCBE has historically relied on its content area experts to validate the content of items during test development, but has recently moved to providing additional content checks by external reviewers (Case, 2005). However, because content validation focuses on items and test specifications, that is, aspects of the instrument or form, use of content validation alone begs the question of the connection between observed test scores and underlying psychological traits. "Considerations of content alone are not sufficient to establish validity even when the test content resembles the trait" (Loevinger, 1957, p. 657) because behavior on test items is always used as a sign of behavior outside of the test situation. Therefore test behavior and the validity of inferences about test behavior are of greater importance than specific test items, which at best are non-reactive sample stimuli used to elicit behavior, and at worst are sources of instrumentation error. Content validation alone is not sufficient, especially when stakes are high.

Due to the lack of a clear external criterion and the limitations of the content approach to test validation, interpretations of scores

on the MBE should be validated from an approach that is focused on constructs. One method used to provide evidence about the causal link between measurement constructs and test scores in the absence of true experimentation is to gather information about the internal structure of a set of test scores to confirm a prediction about the attributes measured by different subscales. A set of test scores may also be compared to scores from multiple other measures to show that relative rankings between examinees converge with different measures of the same attribute and discriminate between measures of less similar attributes as predicted by theory. Empirical confirmation of such complicated patterns of prediction tends to support the theory that the attribute causes the test scores. Researchers may also draw on quasi-experimental evidence about whether groups that are known to be different on the measurement attribute show significant differences in test scores on the average.

Another method of gathering evidence about the causal factors underlying test scores is the method used in the present study: to examine the substantive cognitive processes that drive examinee response behavior. Researchers taking this approach attempt to describe the mental processes that examinees use to generate responses to test items, and to deduce whether those processes are consistent with those predicted by theory. Cognitive processes relevant to task performance can be logically or empirically compared to explicit cognitive models or to more general expectations about task performance to validate a given test interpretation. Analysis of cognitive processes "informs the construct validation of the item set and feeds back to knowledge of what it means to be an expert within the content domain (Leighton,

2005, p.8). This method frequently relies on examinees' verbal reports of their cognitive processes as they complete assessment tasks, as in the present study.

Research Questions

A first problem that the present study addresses is the effect of thinking aloud on participants' performance on cognitively complex multiple-choice questions such as those found on the MBE. If performance is affected by thinking aloud, the use of verbal protocols should be considered reactive with performance. However, the content of the verbal protocols may still be useful for shedding light on response processes associated with different levels of performance, although those performance levels would not be considered typical of the same examinees under non-verbalization conditions.

The study then addresses the following substantive research questions about the MBE through quantitative and qualitative analysis of the verbal protocols: (1) Which mental processes are most frequently used in answering the selected items? (2) What is the relative effect on performance of reasoning from legal principles as compared to general reasoning or use of testwiseness principles? (3) When variance in performance due to similarity of responses to expert problem-solving models is accounted for, do divergent mental processes add to prediction? (4) Do minority ethnic examinees use different cognitive processes or commit different types of errors when answering items in selected MBE content areas as compared to majority ethnic examinees?

CHAPTER 2

REVIEW OF THE LITERATURE

The literature relevant to the present study can be categorized into two broad groups, one focused on the nature of the measurement construct underlying scores on the MBE, and the other focused on the think-aloud method. Therefore the chapter is divided into two sections. The first presents a brief description of the Multistate Bar Examination, then provides a review of the literature relevant to an understanding of the measurement construct, which includes both research specific to the bar exam and viewpoints on the mental processes that characterize legal reasoning. Selected research findings from other fields that are highly relevant to the investigation of mental processes for solving problems in law are summarized. The second section of the chapter focuses on the literature surrounding verbal protocol analysis. In this section, descriptions of the method, research on the validity of the use of the method, and recent examples of the use of the method for test validation are described. The second section closes with a discussion of various approaches to the analysis of data from verbal protocols.

The Multistate Bar Exam and Legal Reasoning

The test used in the present study was the 1998 form of the MBE, which has since been retired. In content outline, length, and administration the 1998 test form is the same as the current form. The exam contains 200 items, each of which presents a fact situation of one or two paragraphs in length. The fact situations are vignettes that

describe two or more parties in a context that raises legal issues. Legal issues addressed in the fact situations on the MBE may be argued under any one of the six tested content areas of the law: contract law, real property, rules of evidence, criminal law, torts, or constitutional law. Items on the test are not grouped by content area, nor is the content area of an item identified for the examinee. Following the fact situation, the test item asks a direct question or poses a problem for the examinee to solve by selecting from one of four options, each of which relates to a different interpretation of key legal principles involved in the issue. Most items are independent, with the exception that on the 1998 test form, two or three items in contract law often shared a reference to a common contract. In these cases, each item had its own short fact situation, which was to be read and interpreted independently, without reference to the other item that related to the same contract.

The test is administered over a six-hour period with a single break. This gives examinees an average of 1.8 minutes to respond to each item. There is no penalty for guessing on the MBE. The MBE is administered twice a year, in the early summer when it is mostly taken by recent law school graduates, and in the winter when a large proportion of examinees are those who have previously failed to pass their state bar.

Previous studies of the MBE have gathered some empirical evidence about the nature of the construct underlying MBE scores. Scores on the MBE do not appear to be a function of specific content area expertise. Empirical data show that examinees proficient in one content area tend to be proficient in all content areas. The average Pearson product-

moment correlation between content area subtotals for the 1998 test was 0.54; disattenuated for unreliability in the subscales, the average correlation coefficient was 0.90. 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, 1993, p. 18). For this reason the NCME does not recommend basing educational or other decisions about individual students on subtest scores. The internal consistency of the total test was high, ($\alpha = 0.90$). High internal consistency may indicate that the test measures a homogeneous construct, although it may be partly an effect of the large number of test items.

Further information from previous research about the nature of the measurement construct of the MBE is somewhat sparse. Klein (1993) cites a 1992 study that looked at the correlation of MBE scores with law school GPA, bar exam essay scores, and LSAT scores. In 13 of the 15 largest law schools in California, the MBE correlated somewhat more strongly with comparatively direct measures of legal skills (LGPA, median $r = .62$, bar exam essay scores, median $r = .64$) than with measures of general reasoning (LSAT, median $r = .58$). 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 further reports a study comparing the

performance of 1st-year California law school students (a group with high LSAT scores and therefore presumptively high general reasoning skills) with the performance of law school graduates. The fact that the graduates did much 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.

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. The discussions one finds about the thinking processes of lawyers are mostly historical and descriptive, or prescriptive. For example, Alexis de Tocqueville, an early commentator on American lawyers, wrote in *Democracy in America* in 1835, "Men who have made the law their special study have learned habits of orderliness from this legal work, a certain taste for formalities, [and] a sort of instinctive love for a logical sequence of ideas" (pp.308). De Tocqueville also credited lawyers with a habit of proceeding slowly, which he contrasted with the general American population's impetuosity.

An exception to the general descriptive tendency is Smith's 1977 study of cognitive styles in law school. Smith rated law school students (n = 782) on four scales: authoritarianism, tolerance of ambiguity, opportunism, and legalism. Scores of law school students were reported to be normally distributed on all four scales; unfortunately, due to the lack of a comparison group, it is unknown if

they were different from those not trained in the law. The only statistically significant correlation found was between grades in a torts course and legalism, which scaled students' orientation from the dogmatic to the pragmatic. However the magnitude of that correlation was very small (r = -.10).

The fact that differences between high legalistic thinking (dogmatism) and low legalistic thinking (pragmatism) do not strongly predict the outcome measure may indicate the hybrid nature of legal reasoning as it is defined today. The tension between dogmatism and pragmatism has been central to the historical evolution of legal reasoning in America. In the late nineteenth century, a tradition of legal formalism developed in the United States, based on the belief that the process of observation and induction would, in time, yield "the discovery of a small number of very general rules. All future cases could be decided by applying these general rules to the facts through the process of deduction" (Vandevelde, 1996, p. 115). The formalist school of thought in law was successfully challenged by a critique based in realism, which "disfavored grand statements of abstract rules …[that] offered a false hope of predictability and obscured the true nature of the policy judgments that were required" (Vandevelde, p. 125). However, realism did not offer a useful theory to replace formalism. Today's mainstream legal reasoning is a modified formalism, combining the mechanical use of rules in cases that are relatively straightforward with a recognition that rules are applied in historical contexts and can change according to public policy needs.

Although modern lawyers use a mixture of dogmatism and pragmatism on a case-by-case basis, basic logic and avoidance of logical fallacies

are still fundamental values. For instance, an author writing for an audience of pre-law and law students states, “In legal argument our major premises must not be based on emotion or instinct” (Aldisert, 1997, p. 35). Advocating an orderly, formal, and logical analytic style similar to that described by de Tocqueville two hundred years earlier, Aldisert warns against the use of fallacious reasoning, including both formal fallacies and informal fallacies. Among the informal fallacies eludicated are the fallacy of irrelevance (missing the point), the fallacy of accident (overlooking exceptions to the general case), and the fallacy of hasty generalization, also known as “jumping to conclusions.” The tendency today to maintain at least a partially traditional and dogmatic approach is further evidenced by the fact that many teachers of law use a method known as IRAC (Issue-Rule-Application-Conclusion) as a heuristic for teaching the structure of legal analysis. The simple nature of this model suggests that for many of those engaged in training future lawyers, the relatively formalistic and mechanical approach of identifying rules and applying them to reach conclusions is viewed as a valuable part of the legal education, perhaps prerequisite to critiquing and modifying dogma based on evolving societal mores.

Synthesizing these diverse views, the construct of legal reasoning may be defined as the application of legal rules and principles to specific cases in light of relevant facts, combined with the ability to reinterpret such rules and principles on a case-by-case basis in the pragmatic light of public policy needs. This type of thinking would be discriminated from thinking processes that were intuitive, illogical, emotional, or creative.

A comparison to similar research in problem-solving in other content areas completes this section. Information about such research is of interest because an analysis of the similarities and differences between problem-solving approaches across domains increases understanding of the generalizability of individual research findings, and adds information about particular problem-solving approaches potentially relevant to the given task environment, in this case the MBE.

Newell and Simon's(1972) work on human problem-solving is an in-depth theoretical discussion against which most current research findings are compared. Their method was to develop theories of behavior "inductively…from close observation of the behavior itself and adapting the theory to differences in task environment and response" (p. 787). Their close observations were frequently aided by subjects' verbal reports of mental processes. One of their important findings was of the importance to successful problem solving of qualities of the specific task environment, which elicits a search in working memory within a problem space that either exists in long-term memory or must be created. Differences in problem-solving performance can be attributed to the adequacy of the problem-solving space in terms of its wealth of declarative and procedural knowledge structures. Some subroutines within a problem space are general; for instance, addition strategies for regrouping generalize to multiple specific addition problems. However, Newell and Simon's information processing model suggests that if a subject's problem space lacks specific subroutines or knowledge structures, developing other solution methods will decrease problem-solving efficiency. Newell and Simon's (1972) conclusions were based in

particular on analyses of problem-solving in chess, where experts were found to have a large "vocabulary" of subroutines or strategies held in long-term memory and available to working memory on presentation of specific stimuli.

Evidence from specific fields of expertise has tended to support Newell and Simon's view. For instance, in the area of medical problem solving, Elstein, Shulman, and Sprafka (1978) found that "heuristic processes were often of secondary important to the students' knowledge of the medical content required to solve the diagnostic cases" (p. 271). In a retrospective on their findings, Elstein, Shulman, and Sprafka (1990) state that "Contrary to our initial expectations, we found that problem-solving expertise varied greatly across cases and was highly dependent on the clinician's mastery of the particular domain" (p. 13). Expert problem solvers did not differ from novices in problem-solving strategies or employment of heuristics so much as in the extensiveness of their repertoire of knowledge and content-specific experience. Perkins and Salomon (1989) summarized other results in the literature supporting the importance for expertise of a large knowledge base of domain-specific patterns and rapid recognition of situations appropriate for the application of such patterns.

Results from other analyses comparing problem-solving approaches between expert and novice problem-solvers differ, however, in that experts have been found to use more hypothetico-deductive general thinking strategies than novices when confronted with complex problems. For instance, in a study of problem solving in physics, Chi, Feltovich, and Glaser (1981) concluded that "Novices' schemata may be characterized as containing sufficiently elaborate declarative

knowledge about the physical configurations of a potential problem, but lacking abstracted solution methods" (p. 151). These differences in solution methods between experts and novices were supported in an experimental study by Schoenfeld and Hermann (1982), which suggested that experts perceived problems in terms of deep structure while surface structure was the primary criterion used by novices. Clement's (1989) research on the thinking processes that lead to new hypothesis formulation in science also suggests that experts rely on an eclectic blend of deep structure strategies including analogy, association, and transformation to solve novel problems.

The differences in these two sets of results, one holding content-specific declarative and procedural knowledge routines as the primary feature that distinguishes experts from novices, and the other holding declarative knowledge relatively constant while solution strategies varied between the two groups, may relate to the way experts and novices are defined in the specific studies, or to aspects of the specific tasks. After all, the distinction between novices and expert problem-solvers is an artificial binary one representing an underlying continuum of expertise, and tasks can range from relatively routine and restricted problems to "thorny" problems required novel approaches and creative solution strategies, such as those studied by Clement (1989). The two strands of research findings are elegantly and usefully synthesized by Perkins and Salomon (1989) who answered the question "Are cognitive skills context-bound" with the response, "Yes and no", claiming that while cognitive skills are context-dependent, general heuristics "help when experts face atypical problems in a domain" (p. 23).

For the study of problem-solving approaches likely to be of greatest approaches for performance on a professional licensure examination such as the MBE, it therefore remains to define the level of expertise expected and the type of problem. Fairly typical problems intended to tap into minimal expertise would be expected to draw primarily on declarative or basic subroutines of procedural knowledge, while novel problems intended to tap into high levels of expertise would be expected to elicit deep structural insights and more general problem-solving strategies based on analogical reasoning and more creative approaches. In the case of the MBE, the purpose of the test is not to identify examinees with high levels of expertise and deep insights, but to identify and bar from professional practice those who do not possess basic knowledge by assessment of critical competencies. "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). Therefore if the test is appropriately targeted, it would be expected that high performers on the MBE would be those with greater knowledge of legal principles, not necessarily those with deep structural insight into the law. If the construct of legal reasoning contains both the components of dogmatic application of legal principles and need-based reinterpretation of such principles, the latter ability would be expected to lie outside the scope of measurement of the MBE.

The Method of Verbal Protocol Analysis

As a data-collection method, the study draws on verbal protocol techniques described by Ericsson and Simon (1993). Verbal protocols, commonly known as "think-alouds," have 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 application of think-aloud protocols to assess the validity of test interpretations is far from new. In a 1950 study, Bloom and Broder compared think-aloud protocols of high- and low-scoring examinees on a test of reasoning (Bloom & Broder, 1950). As they describe it, "In order to secure data on the thoughts, feelings, and methods used by students when attacking this group of problems, an attempt was made to get each student to 'think aloud' as he worked the problems" (p. 8). Understanding of the mental processes students use in responding to test items is essential for test validation, because "the reactions of the student in the context of a test situation would appear to represent the basic kinds of information the test constructor must have if he is to do more than improve the statistical picture of a test. If he is to gauge the value of the problem or the test exercise, he must have insight into the thoughts, feelings, and mental operations of the student when confronted with a specific problem. Until he has such insight, the test constructor can only proceed by trial and error, making many test exercises and retaining those which have appropriate statistical properties" (p. 66).

Drawing on this reasoning, think-aloud protocols have been recently used to help validate test scores in several studies of hands-on science assessments (Yepes-Baraya, 1996; Hamilton, 1994), online

problem solving assessments (Chung et al., 2002), and as a way to gather evidence about the validity of scores from the constructed-response and multiple-choice format tests used in the National Education Longitudinal Study of 1998 (Hamilton et al., 1997). Verbal protocol analysis has also been used to study the cognitive processes of raters in portfolio assessment (Heller, Sheingold, & Myford, 1998). These studies, some of which are described in detail below, illustrate that measurement experts continue to recognize the potential of thinking aloud for test validation.

However, accompanying that sense of potential has been lack of confidence in the method. Cronbach (1971) observed that, while asking students to work multiple-choice problems out loud could add insight to amplify the meaning of the measurement construct, there were at the time few reported examples of such validation, due to "difficulties in using the oral-response technique" (p. 474). In the Bloom and Broder (1950) study, for example, lack of a control group or repeated measure made it impossible to assess the effect of thinking aloud on performance. Also, the same method was not used with all participants in the study, as some subjects (who spoke too quickly for note-taking or with little detail about their processes) were probed for retrospective reports on their processes and others were not. Bloom and Broder admit considerable variation within the study in interviewing techniques. More recent research has revealed that interviewing techniques may substantially affect primary problem-solving processes (Ericsson & Simon, 1993).

The methodological difficulties that surround gathering and analyzing verbal protocols have been a subject of debate in virtually

all the fields that use the method. They have 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. The principal modern arguments against the use of people's self-reports of mental processes are given by Nisbett and Wilson (1977). In an often-cited article, 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 where research participants have failed to report, inaccurately reported, or even denied recognition of mental processes which had been induced in them by the experimenter.

Two examples from Nisbett and Wilson (1977) will serve to demonstrate their argument that research participants are frequently unable to articulate their own mental processes. In one experiment demonstrating the halo effect, students were shown an interview with a college teacher who spoke accented English. Researchers experimentally manipulated the apparent warmth of the teacher's personality. The students then rated the teacher on likability as well as on physical attributes that were invariant across experimental conditions, such as appearance and accent. Students who saw the warm personality liked the teacher and also his physical attributes better to a significant degree than students who saw the cold personality, demonstrating the halo effect. Demonstrating the erroneous nature of verbal explanations of mental processes, students in both conditions denied that their liking

of the teacher's personality had influenced their liking of the attributes. Additionally, students who saw and disliked the cold personality actually reversed the causal order of the relationship between their dislike of the teacher's attributes and his personality, stating that their dislike of his accent and appearance had actually caused them to dislike his personality.

In another demonstration experiment, researchers asked passersby to choose the best quality pair of nylon stockings from a selection of four identical pairs. 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. However, when asked to report the reason behind their choices verbally, the passersby never included a mention of the socks' position. In fact, when they were asked if position might have had an effect on their choice of socks, they all denied the possibility.

These two experiments and others demonstrate that at least some mental processes underlying people's preferences and choices are not available to the conscious mind. They therefore form the basis for an argument against the use of verbal reports of thinking, because if people are unconscious of or are regularly mistaken about their own mental processes, their verbalizations will be useless or misleading. Nisbett and Wilson (1977) generalize about their results to describe characteristics that will tend to make verbal reports of mental processes inaccurate. According to Nisbett and Wilson (1977), verbal reports will most tend to inaccuracy when the verbalization is removed in time from the mental process it relates to. Also they will likely be inaccurate when the cause of behavior seems implausible or irrelevant,

as in the case with “mechanics of judgment” effects such as position and order effects, context effects, the effects of nonverbal behavior, and large effects from small causes.

The opposing evidence in support of the use of the analysis of verbal protocols has been summarized by Ericsson and Simon. Ericsson and Simon (1993) recommend analyses based on what they call Level 2 verbalizations, which are concurrent verbalizations that are intended to communicate thinking processes based on information heeded in short-term memory directly without intentional introspection or reflection. They assign to this level “verbalizations that do not bring new information into the focus of the subject’s attention, but only explicate or label information that is held in a compressed internal format or in an encoding that is not isomorphic with language” (p. 79). According to Ericsson and Simon (1993), Level 2 verbalizations tend to increase response time but do not affect performance: “Since explication or recoding requires processing time for the subject but does not replace other processing involved in the task performance, a subject who is verbalizing at this second level can be expected to take more time for the task than one who is not verbalizing. However, we would hypothesize that such recoding does not change the structure of the process for performing the main task” (p. 79).

Ericsson and Simon 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. Some types of mental processes, however, are held to be inherently unavailable for concurrent verbalization. Retrieval of highly automated information and simple

recognition are mental processes that apparently leave little or no trace in short-term memory and are therefore not accessed during verbalization. Also, verbalizing may interfere or alter some kinds of pictorial and spatial thinking where information is not verbally encoded.

Ericsson and Simon (1993) contrast concurrent verbalization of appropriate cognitive tasks with Level 1 verbalizations, which are personal forms of thinking aloud, not intended for communication purposes, and therefore highly idiomatic and idiosyncratic, and with Level 3 verbalizations, which are introspective and retrospective. According to the information processing model adopted by Ericsson and Simon, when awareness is assessed though probing after an experiment as in a Level 3 verbalization, the relevant heeded information is no longer available in short-term memory and must be intentionally retrieved from long-term memory. "Memory retrieval is [sic] fallable, sometimes causing access to other related, but inappropriate, information. Further, what information can be recalled depends on what cues and probes are provided. Hence the completeness of the information retrieved will vary with the probing procedures" (p. 140). According to Ericsson and Simon's model, most if not all of the studies described by Nisbett and Wilson (1977) where verbal reports provided erroneous explanations would fall into the Level 3 category.

The potential issues about the validity of verbal protocols can be categorized into problems of two types: reactivity and veridicality (Russo, Johnson, & Stephens, 1989). Reactivity occurs if verbalization changes the primary process, rendering the verbal protocols invalid for making inferences about the primary process. The primary process may be

changed in outcome or in response time. Verbalization has been very commonly found to lengthen response time; however, this is not considered a serious effect when time is not the object of inference. Possible changes in performance due to verbalization are considered more serious sources of invalidation. Russo, Johnson, and Stephens (1989) conducted extensive experiments on reactivity in concurrent verbalizations, with four dissimilar tasks of 45 problems each administered to 24 subjects, randomly assigned to different experimental conditions: silent (control), concurrent verbalization, response-cued, stimulus-cued, and prompted. The predicted longer response times in the verbalization conditions were found for all tasks. The experimenters also found significant differences in accuracy between verbalization and nonverbalization conditions in some of their tasks, with concurrent verbalization associated with increased accuracy compared to the control in a task involving choosing between two gambles, and decreased accuracy in a task involving adding three-digit numbers. The decreased accuracy in adding three-digit numbers was theoretically based on competition for resources in working memory. The increase in accuracy for the gambling task ran counter to prediction and suggested to the researchers a possible motivational shift among subjects in the concurrent verbalization condition. These experiments therefore yielded mixed results about reactivity with concurrent verbalization, suggesting that nonreactivity might be possible with some tasks but should not be assumed.

In another fully randomized experiment on the effect of eliciting verbal reports of thinking, Norris (1990) administered a multiple-choice test of critical thinking to 343 high school students who had

been randomly assigned to five groups. In four of the groups, verbal reports of thinking were elicited. One of the verbalization groups used concurrent verbalization or think-aloud instructions, while those students in the other three conditions were asked to explain their reasoning after they arrived at each problem solution, or were asked specific probing questions retrospectively. No statistically significant effects on test performance for verbalizing compared to silent responding were found. This provides evidence that the elicitation of verbal reports of thinking on this type of higher-order multiple-choice test is not reactive with performance; however, it does not provide any evidence as to whether such verbal reports provide enough information to be useful for test validation.

Veridicality refers to the degree to which subjects are able to report their own thinking accurately, absent such errors in self-description as those described by Nisbett and Wilson. Veridicality is reduced both when people omit relevant thinking in their verbalizations, and when they report thoughts that have not occurred. The Russo, Johnson, and Stephens (1989) experimental data also permitted the comparison of concurrent and retrospective verbalizations of the same task performance. Comparison of the two sets of reports suggested that the retrospective verbalizations had lower veridicality than the concurrent verbalization protocols due both to forgetting (which resulted in errors of omission) and fabrication (which resulted in errors of commission). Leighton (2005) argues that distortion or lack of veridicality are primarily associated with the sort of attitudinal retrospective reporting reviewed by Nisbett and Wilson

rather than with concurrent reports of problem-solving on cognitive tasks.

Even Nisbett and Wilson (1977), while arguing against the use of verbal reports in general, recognized that verbal reports could be accurate in some contexts, particularly when they are elicited quite close in time to the application of the stimulus. They state that "reports will be accurate when influential stimuli are (a) available and (b) plausible causes of the response, and when (c) few or no plausible but noninfluential factors are available" (p. 253).

As evidence in favor of the validity of some verbal reports of mental processes, Nisbett and Wilson (1977) report on literature where clinical psychologists have been asked to state the subjective weights of various factors in their diagnostic decision-making, and then the subjective weights are compared to objective weights derived by regression. The existence of correspondence between subjective and objective weights demonstrates an instance where people apparently are at least to some degree accurate in reporting the effects of stimuli on their mental processes. Nisbett and Wilson (1977) account for this "lonely outcropping" of accuracy in verbal reporting thus: "[I]n general, we may say that people will be accurate in reports about the causes of their behavior and evaluations wherever the culture, or a subculture, specifies clearly what stimuli should produce which responses, and especially where there is continuing feedback from the culture or subculture concerning the extent to which the individual is following the prescribed rules for input and output" (p. 254).

In fact, according to a more recent review of the arguments about verbal reports (Pressley & Afflerbach, 1995), 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 only information held in short-term memory, 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 prone to poor veridicality. According to Pressley and Afferblach (1995), Ericsson and Simon provide the "state-of-the-science" approach to how to do protocol analyses, and a definition of what constitutes an adequate verbal report of a subject's cognitive processing and response.

Even Ericsson and Simon (1993) concede that concurrently elicited verbal reports are not likely to be complete, and that some mental processes such as the processing of specific cues prior to the spark of recognition may not be available to conscious processing. However, they argue that verbal reports need not be exhaustive to be enlightening. Evidence that verbal reports of thinking can be useful for explaining behavior and predicting performance argues in favor of at least partial veridicality. For instance, researchers in the field of decision theory often adopt a process-tracing approach in which evidence from think-aloud protocols is used to reveal types of strategies, information preferences, and thought sequences relevant to decision-making. Harte, Westenberg, and Van Someren (1994) found 23 studies published between 1976 and 1993 that used think-aloud protocols obtained from concurrent verbalizations to study individual decision-making on well-defined problems. Some of the studies used results from analysis of the verbal

protocols to build theory, but others had a more confirmatory, theory-testing purpose, demonstrating the usefulness of the method.

A few examples will illustrate the usefulness of verbal protocol analysis in educational measurement. Yepes-Baraya (1996) administered two blocks of items from the 1993 NAEP science field test to 16 eighth-grade students who had been selected from different percentile ranges (low, middle, and high) on a standardized achievement test. Within two weeks after completing the assessment, students were readministered the items and asked to think aloud while responding. Students were also questioned about their thinking after completing each item. Although the method used in that study clearly did not meet Ericsson and Simon's criteria for Level 2 concurrent verbalization, Yepes-Baraya reported that "the combination of the think alouds and concurrent interviews following the administration of the assessment proved to be an effective way to better understand students' cognitive processes as they worked on the NAEP science field test" (p. 14). His results showed that the NAEP science field test elicits mostly thinking processes that are relevant to the intended measurement constructs: science knowledge structure, reasoning, hypothesis formulation, and hypothesis testing. Yepes-Baraya also was able to detect through analysis of think-aloud protocols assessment items for which difficulty was related to reading comprehension, which was considered to be construct-irrelevant. Because the think-aloud study was conducted as part of field testing, those items whose variability was influenced by reading comprehension could be modified or eliminated based on the results.

Hamilton, Nussbaum, and Snow (1997) used think-aloud techniques to increase their understanding of subscores on a science achievement

test derived by factor analysis from the NELS data set. Forty-one high school students were asked to think-aloud while taking 16 multiple-choice items from the science achievement test as well as three constructed-response items. They reported that the students' verbal responses helped to support their understanding of the subscores and define the constructs underlying subscores more clearly. "Interviews also helped to identify unanticipated cognitive processes employed by test takers on constructed-response and performance tasks." This study combined concurrent verbalizations with prompted retrospective interviews.

Katz, Bennett, and Berger (2000) used think-aloud techniques to study the effect of response format on difficulty of SAT math items among 55 high school students at different ability levels. They sought to investigate whether problem-solving strategy mediated the effect of format on difficulty, that is, if examinees chose different strategies depending on item format that resulted in different item difficulties. For instance, in multiple-choice problems students might be expected to use more estimation strategies than on constructed response items. They reported that most research on the relationship between solution strategy and response format effects had inferred solution processes from item statistics, and noted that "the results of processing may reveal little about how that processing actually occurred" (p. 39). The think-aloud method allowed the examination of solution strategies when response format was varied at a different level of inference. Format effects were found in one of two sets of items, but analysis of item-level solution strategies from the verbal protocols did not find a

relationship between type of strategy (estimation versus derivation) and format.

As a final example of the use of verbal protocol analysis in educational measurement, Bond (1990) sought to investigate the usefulness of an information-processing approach in explaining the performance of minority students on the SAT-M, and additionally, "to test the notion that poor performance is primarily a function of a lack of procedural skills and not the absence of an adequate declarative knowledge base" (p. 95). His results, based on concurrent verbalizations of 34 high school students, showed that poor performers had inert, unintegrated knowledge and lacked automaticity, so that they spent much of their testing time consumed in solving routine subproblems, and "often lost track of the original problem and had to reread the problem often to retain the ultimate goal in working memory" (p. 105). Comparing the verbal protocols of black and white students who performed well in math courses but had SAT math scores below 450, Bond (1990) found their problem-solving strategies to be indistinguishable.

A separate issue with the use of verbal reports of thinking is how to analyze the rich data from the verbal protocols. The various methods used can be classified roughly into three groups, with many researchers using multiple methods, as in the present study. The first approach does not attempt to quantify the data or to apply to it any statistical analyses. In this approach the researcher neither makes nor tests hypotheses, but attempts to discover and describe the underlying structure of the cognitive processes encountered in each verbal protocol through the use of a combination of reason and intuition. Then

the individual structures of thought are compared, again using reason and intuition, to articulate generalities across individuals that characterize the problem or object of thinking (Aanstoos, 1982). In test validation studies, this approach is exemplified in Hamilton's (1994) study of hands-on science assessments that used concurrent think-alouds and interviews to explore the cognitive demands of cognitively-complex performance-based assessment tasks among a small number of 6th grade students without any attempt at quantification.

A second approach uses iterative methods to categorize elements from the verbal protocols into classes derived either from theory or data. As described by Chi (1997), "the qualitative data is examined for impressions and trends, methods of coding are developed to capture those impressions, and the codings can then be analyzed quantitatively" (p. 7). This approach can be considered exploratory according to the extent to which the basis for reducing and coding the data is driven by the nature of the contents of the verbal protocols themselves. A variety of analyses can be performed, for instance relating use of strategies of interest to performance, comparing experts to novices for between-group differences in use of strategies, or even testing for the statistical reliability of coded protocols that have been graphically depicted in some form. Elstein, Shulman, and Sprafka (1978) essentially took this approach in their study of medical reasoning with a specific, fixed-order set of problems. They developed a set of variables or categories to look for in the verbal protocols during pilot testing which they then used as predictors in a series of repeated measures analyses of variance.

A third approach derives models for problem solving from theory and task analysis, sometimes supported by pilot testing. The models typically include not only properties of the task but sequence of task completion or procedures. Computer models can then be constructed which can be statistically compared to the evidence in the protocols, as described by Ericsson and Simon (1993, p. 195). Alternatively, similarity between the models and the protocols can be rated by judges. In some cases the construction of a valid computer model for the problem-solving task is the goal of this type of research; however, Van Someren, Barnard, and Sandberg (1994) describe a method for quantifying fit between a model and verbal protocols without relying on computer simulation (pp. 131-133). The theory-driven, confirmatory nature of this method differentiates it from the previous two methods described.

Summary

The relevant literature regarding the MBE has been summarized in this chapter to describe the test, show the limitations of existing research on the construct measured by the MBE, establish the evidence about a performance gap between white and nonwhite test-takers on the MBE, and narrow and refine the definition of the measurement construct. Based on the literature review, the construct of legal reasoning has been defined as the application of legal rules and principles to specific cases in light of relevant facts, combined with the ability to reinterpret such rules and principles on a case-by-case basis in the pragmatic light of public policy needs. The construct is specifically distinguished from intuition, emotion, fallacious reasoning, and creativity. Research on problem-solving in other domains was presented

to demonstrate the importance of a strong repertory of context-dependent declarative and procedural knowledge for performance in relatively routine, restricted problem-solving, when high performance is defined as minimal competency rather than true expertise.

The chapter then presented evidence in support of the type of verbal report of thinking used in the present study, as well as important evidence that for certain types of thinking and in certain circumstances, verbal reports are frequently erroneous. A number of recent studies that profitably used such reports, especially in educational measurement, were described. The method used in the present study is expected to provide relatively complete, veridical, and nonreactive data because the methodology used meets the requirements for a Level 2 concurrent verbalization as described by Ericsson and Simon (1993), and does not involve responses to particularly problematic stimuli such as simple recognition, or pictorial or spatial thinking. Furthermore, the type of problems attempted by participants in the present study would tend to be supported by Nisbett and Wilson (1977) as involving a specific culture with clearly prescribed rules for input and output.

CHAPTER 3

METHODS

Participants

Participants were recruited from the Rogers College of Law at the University of Arizona in June of 2004 to participate in the think-aloud study. Rogers College of Law is one of two large law schools affiliated with public universities in Arizona. It is ranked 41 by U.S. News and World Report (2005) in their rating of U.S. law schools based on a weighted average of measures of quality (peer and professional assessments), selectivity, placement success, and faculty resources. Comparisons between the student characteristics of Rogers College of Law and other institutions are displayed in Table 1. Rogers College of Law is ranked near the middle of the top 100 U.S. law schools by U.S. News and World Report, as are the University of Florida's Levin College of Law, and the law schools of the University of Maryland and Arizona State University. It is somewhat less selective than the University of California at Los Angeles, and somewhat more selective than the University of Arkansas. In 2004, 80.3% of Rogers College of Law graduates taking the Arizona Bar Exam passed, compared to 66.5% of Arizona examinees overall.

Eligibility was based on whether prospective participants had completed courses in the relevant subject areas (offered in the first year), planned to take the MBE, and had not already been exposed to the retired test items. Preference was given to prospective participants who were actively preparing for the MBE and therefore had strong

Table 1[1]: Comparisons among law schools

Law School	Median LSAT	Undergraduate GPA (25^{th}-75^{th} %iles)	Acceptance rate
University of Arizona Rogers College of Law	160	3.24-3.67	23%
Arizona State University	156	3.07-3.62	24%
University of Arkansas (Little Rock)	154	3.05-3.66	54%
University of California (Los Angeles)	164	3.47-3.77	19%
University of Florida Levin College of Law	157	3.26-3.76	27%
University of Maryland	156	3.15-3.66	40%

[1] Based on information provided by U.S. News and World Report (2005).

internal motivation to do well on the items. Recruitment was conducted through flyers, email announcements to the Rogers College of Law listserves, and announcements in test preparation classes. A combination of financial reimbursement for time ($10 per hour) and opportunity to receive feedback on performance on MBE-type items was offered as incentive for participation.

The sample of convenience was composed of twenty-five law school graduates who were actively engaged in preparing for the June 2004 MBE administration at the Rogers College of Law. The description of the sample according to sex and ethnic background is given in Table 2. Three of the participants were non-native English speakers who also belonged to minority ethnic groups; two of them were Asians and one Latino. Twenty-two of the participants were recent (May) graduates of the Rogers College of Law; the remaining three participants had attended law school outside Arizona but were preparing to take the Arizona bar exam. The participants were representative of Rogers College of Law students in gender makeup and ethnic diversity. The 2004 entering class at Rogers College of Law was composed of 50% female and 32% minority students; the sample was composed of 52% female and 36% minority students.

Materials

Retired test items from the 1998 MBE were obtained in electronic format from the 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). Once

Table 2: Sample characteristics

	Sex		
Ethnic background	Male	Female	Total
Asian	2	0	2
Black	3	1	4
Latino	0	3	3
White	7	9	16
Total	12	13	25

the content areas were decided, nine items were selected from the 1998 exam for use in the think-aloud study. Three items in each content area were chosen from the pool of 33 to 34 items per subtest. In addition, two items were selected to be used as practice, to help participants get accustomed to verbalizing their thoughts. Items were chosen largely based on their discrimination index and difficulty. First, items were identified in the selected content areas that had larger than average point-biserial correlations in the 1998 test administration. From this field items with difficulties between 35 and 65 were selected, to maximize variability in participant responses. From this small remaining pool, three items in each content area were selected that had the highest point-biserial correlations in the 1998 examinee group, plus two more, for the practice items. Then 13 additional items were selected from each of the three content areas for participants to answer silently as a repeated measure of exam performance. The average difficulty and discrimination indices for the set of nine verbalization items and the set of 39 silent response items from the 1998 test administration is provided in Table 3, along with the total 1998 MBE average difficulty and discrimination, for comparison purposes.

Procedure

Prior to conducting the study, two pilot think-alouds were conducted to check the procedure and equipment. The equipment consisted of a small audiotape recorder and a clip-on microphone. A research assistant participated in one of the pilot think-alouds and also received training on the materials. Five of the participants in the

Table 3: 1998 MBE and item subset item indices

	Average population difficulty[2]	Average population discrimination[3]
Verbalization items (9)	.52	.33
Silent items (39)	.69	.25
Total MBE (200)	.63	.21

[2] The proportion correct from the June 1998 MBE test population, N = 43,541

[3] The point-biserial correlation between the correct response and MBE total score, June 1998 MBE test population.

study were interviewed by the research assistant, a female undergraduate psychology major.

The first subset of nine MBE items was administered to each participant individually. Participants were asked to think aloud about their response strategies while being audiotaped, following a standard protocol that met the requirements for a Level 2 verbalization as described above. Participants were instructed to verbalize all their thoughts without attempting to reflect back upon previous thoughts or to edit their thinking in any way. Participants were reassured that all thoughts as they occurred 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 attempt to provide a full description of their mental processes as they attempted to answer the test items without regard to time. Before asking the participant to sign the consent form, the researcher or research assistant described the purpose and method of the study, and checked for the participant's understanding and comfort.

Once informed consent was obtained and the survey completed, the researcher or research assistant read a prepared script to inform participants about how to verbalize, as follows: "For the purposes of this study, verbalize or think aloud as you read each item and as you select the best response. Do not attempt to reflect on your own performance; simply talk out loud as you think, stating all the mental steps you go through. Don't hesitate to express rethinking, uncertainty, or revisions to your first impressions. There are no 'correct' or 'incorrect' thoughts; all your ideas, strategies, and processes are of interest. We encourage you to speak freely about the

thinking you use to answer these questions. There is no time limit on the think-aloud items." These instructions were given twice, before the warm-up items and before the main verbalization items, to encourage unrestricted verbalization free of any censorship process that might intervene between awareness of short-term memory content and vocalization, and to discourage introspection and explaining. It was considered especially 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. 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.

Two warmup items were administered to make sure participants understood the method and were able to verbalize comfortably. The warm-up items were used as an opportunity for the interviewer to encourage reluctant verbalizers to think aloud more often, and to reassure participants that all their verbalizations were appropriate and to ignore time constraints. After participants responded to the warm-up items, the interviewer asked them if they felt comfortable moving forward with the rest of the items. Once the warm-up items were completed, the interviewer did not further interfere with participants' processes through prompting or questioning, except very minimally if it was needed to remind a participant to continue to verbalize or to repeat something. In such cases, care was taken to ensure that interference was kept to a minimum. These methods for verbalization instructions, warm-ups, and reminders were aligned with recommendations

by Ericsson and Simon (1993) and Van Someren, Barnard, and Sandberg (1994).

After the administration of the verbalization 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 3 hours per 100 items or 1.8 minutes per item. This part of the study measured the participant's test performance under conditions that simulated to the extent possible actual MBE testing.

Following completion of the verbalized and silent items, participants were asked to reflect on the effect of verbalizing on their performance. The participants were asked to categorize the effect of verbalizing on their ability to respond to the test items as helpful, interfering, or having no effect.

After all items and questions had been completed, the interviewers offered feedback to participants. Participants were told the number of items from both the verbalized and nonverbalized sections they had answered correctly, and the correct answer to each item. The information was broken down by verbalization condition and by content area. For the verbalization items, the interviewers offered explanations for the correct responses based on the expert problem-solving models.

The researcher found through discussion with the research asssistant as well as review of the verbalization transcripts that the majority of participants appeared to feel comfortable verbalizing. Most participants used both warm-up items, but a few participants stated

that they did not need both items to warm up to the process. Only two of the 25 participants referred at any time while they were thinking aloud to awareness of the interviewer or the think-aloud process (in both cases, stating that they felt awkward). Reminders to verbalize were very seldom required.

Preparing the Verbal Reports for Analysis

After transcription, the verbal protocols were divided into segments. A detailed discussed with examples of different approaches to the segmentation process is provided by Chi (1997). Complete verbal protocols may be divided at various points, "revealing units of varying grain sizes, such as a proportion, a sentence, an idea, a reasoning chain, a paragraph, an interchange as in a conversational dialogue, or an episode" (p. 9). Katz's (2000) study of format effects on item difficulty is a recent example of work using verbal records that adopted a coarse grain approach, classifying each subject's solution strategy on each problem holistically as traditional or nontraditional on the grounds that "fine-grain analyses are beyond what is necessary to distinguish between the broad classes of solution strategy" (p. 47). 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 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. 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 two ways. The first approach to coding the segmented item responses was based on the 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 which are used to map the verbal protocols. Preliminary categories used for coding the protocols included the following: gathering facts, thinking with reference to specific cases, thinking with reference to legal principles, hypothesis development, evaluating options, guessing, and non-cognitive 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 sufficient system for coding was developed, the rest of the segmented transcripts were coded. 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.

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…"

Note that in the two examples above, "proximate cause" and "consequential damages" are both terms in American jurisprudence used to refer to specific rules and principles. They also were either explicitly named as relevant principles by the legal experts consulted during development of the coding system, or listed as relevant principles for study in the test preparation materials provided by the NCBE. Thus segments categorized under this heading could be clearly identified as drawing on legal principles, even though they might do so in quite different ways, as in these examples. In the first case an element of the fact situation is merely identified as being a proximate cause. More sophisticated analysis is brought to bear in Example B, where the participant makes a judgment about the applicability of a legal principle to the given fact situation from which it has been inferred.

4) Making decisions about options: Segments in which participants articulated decisions about choices such as to eliminate or to 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…"

5) 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…"

6) 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…"

Note that in some cases, such as Example B, this is almost a special case of test-strategizing, as it results in elimination of response options. It is distinguished from test-strategizing in that the elimination behavior results not

from content about the options themselves (which makes them seem false or irrelevant and so liable to be eliminated), but from a conclusion drawn from the fact situation without reference to the content of the options.

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…"

   Note that in Example A, the possibly legalistic term "zone" is given in the fact situation, so that no non-given legal principles are being inferred. Although there may be an unarticulated legal principle underlying the participant's rationale for the utterance, the articulated response only demonstrates an inference about the propriety of zoning based on the intuitive sense of "sounding right," not based on articulated legal principle.

8) Nonsolution-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 non-cognitive verbalizations, meta-cognitive 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…"

The preliminary category of thinking with reference to specific cases was considered a part of using legal principles, due to its very low incidence (3 occurrences over approximately 4,000 coded segments). With that addition, legal principles were those listed by the subject matter outlines provided by the 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.

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, errors about legal principles. These types of errors are very similar to those analyzed by Elstein, Shulman, and Sprafka (1978) which included misinterpretation of data cues, inaccurate hypothesis generation, and errors in evaluation of hypotheses. They are also comparable to the error categories used in Bond's (1990) analysis of quantitative reasoning, which included errors of basic fact and errors of integration. Bond did not analyze his participants' responses in terms of early hypothesis formation, did include errors in use of drawing as an aid, which was a category not pertinent for this study. The error types captured with these categories are all errors of commission; errors of omission are difficult to measure, but are captured to some extent in this study in

the second coding approach described below, where responses are compared to expert, "full" models of problem-solving.

Twenty percent of the transcripts were coded by a second rater to assess inter-rater agreement on coding. The second rater was a female doctoral student in Educational Psychology. Transcripts were systematically selected for the inter-rater 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 inter-rater agreement was about 70% averaged across items and persons, so that the rating system appeared 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 these problems were diagnosed through close examination of the categories where disagreement was high, the raters met for further training. Using the NCBE content outlines and the expert models, 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 inter-rater agreement analysis following this second iteration averaged across all items and all processes are given in

Table 4, as well as agreement on specific key processes. Percent agreement and kappa coefficients were also calculated at the individual item level. Inter-rater agreement did not vary noticeably among items; the lowest kappa coefficient obtained for any item on a key process was 0.84. Based on these results, inter-rater agreement was 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.

In the second coding approach, item responses were rated according to adherence to exemplary problem-solving models derived from task analyses. To create models of expected item-responding processes for each item, test preparation materials and test specifications were initially consulted. These sources yielded general information in the form of subject matter outlines, but not information specific to individual test items. Then law experts were consulted to review the test items. Experts were identified from the faculty of Rogers College of Law at the University of Arizona based on their teaching duties and stated areas of expertise as found on the departmental website, and were asked to participate in the task analysis on a volunteer basis. Law experts were asked to discuss the processes an exemplary test-taker would be expected to use to arrive at the correct response, as well as the considerations that would lead a knowledgeable examinee to rule out the incorrect options. Two experts in each content area were consulted to complete this task analysis, and their responses were pooled to make as complete a model as possible. The completed task models summarized the key concepts needed for correct interpretation of the legal issues in each given fact situation, understanding of the correct application

Table 4: Inter-rater agreement on coding of cognitive processes

	Kappa	agreement
All processes[4]	.87	90%
Inferring based on legal principles	.86	95%
Rehearsing cues	.88	96%
Inferring beyond givens and principles	.77	98%
Drawing early conclusions	.98	100%
Test-strategizing	.87	98%

[4] Including Decision-making, Nonsolution-oriented thinking, Classifying types of problems, and Other (not categorizable)

of legal principles in the keyed response, and understanding of the flawed or incomplete use of legal principles in the distractors. The following example shows an item from the content area of Constitutional Law, followed by the exemplary problem-solving model, which in this case has eleven elements:

Item:

Company wanted to expand the size of the building it owned that housed Company's supermarket by adding space for a coffeehouse. Company's building was located in the center of five acres of land owned by Company and devoted wholly to parking for its supermarket customers.

City officials refused to grant a required building permit for the coffeehouse addition unless Company established in its store a child care center that would take up space at least equal to the size of the proposed coffeehouse addition, which was to be 20% of the existing building. This action of City officials was authorized by provisions of the applicable zoning ordinance.

In a suit filed in state court against appropriate officials of City, Company challenged this child care center requirement solely on constitutional grounds. The lower court upheld the requirement even though City officials presented no evidence and made no findings to justify it other than a general assertion that there was a shortage of child care facilities in City. Company appealed.

The court hearing the appeal should hold that the requirement imposed by City on the issuance of this building permit is

(A) constitutional, because the burden was on Company to demonstrate that there was no rational relationship between this requirement and a legitimate governmental interest, and Company could not do so because the requirement is reasonably related to improving the lives of families and children residing in City.
(B) constitutional, because the burden was on Company to demonstrate that this requirement was not necessary to vindicate a compelling governmental interest, and Company could not do so on these facts.
(C) unconstitutional, because the burden was on City to demonstrate that this requirement was necessary to vindicate a compelling governmental interest, and City failed to meet its burden under that standard.
(D) unconstitutional, because the burden was on City to demonstrate a rough proportionality between this requirement

and the impact of Company's proposed action on the community, and City failed to do so.

Model Response Elements:

Category Classification:

1. Constitutional Law

Observation of major concepts relevant to understanding fact situation:

2. City limits building permit on Company’s property
3. Issue involves takings issue through zoning regulations
4. City did not justify action through evidence

Option analysis:

5. A—Rational basis test is incorrect standard of review for takings
6. A—Placing the burden on Company is correct for rational basis test
7. B—Not an issue of compelling state interest which requires strict scrutiny standard of review and is reserved for fundamentals
8. B—Placing the burden on Company is incorrect for strict scrutiny and/or takings issue
9. C—Not an issue of compelling state interest which requires strict scrutiny and is reserved for fundamentals
10. D—(CORRECT): Burden of proof is correctly placed on City
11. D—(CORRECT): Level of review is correct (proportionality test is part of intermediate standard of review, higher than rational basis and below strict scrutiny).

Legal terms used in the model such as “level of review” and “strict scrutiny” were discussed between the raters prior to rating the items until a shared understanding was reached. Because participants had not been asked to explain their thinking, when their verbalizations related to legal principles they tended to use language similar to the language in the model, which the rater merely needed to recognize. Thus no deep understanding into legal complexities was required from the raters.

Each participant’s transcribed and segmented verbalization of his or her solution process on each item was rated according to how many

unique components of the exemplary model were found in the response. The count of unique concepts in the model that were contained in a participant's response was converted to a proportion, because the total number of key concepts in each item's model varied slightly. A variable was thus created that represented the proportion of the expert problem-solving model matched by each participant's response on each item. This variable captured the extent to which the participant's response referred to or omitted the most important cues from the fact situation and legal principles, and was thus an indirect measure of errors of omission.

All attempts were made to prevent biases in coding due to the correctness of a participant's response. Verbalizations about response choices had been separated from substantive reasoning in the segmentation process, and attempts were made to ignore such response-choice segments in the model-matching coding process. A choice was made, however, to allow raters to see all the segments including the participant's ultimate response choice (when verbalized) rather than edit the response. This was because of the nature of the verbalizations; participants would frequently refer in a segment to a principle or fact element verbalized previously with a phrase like "what I was saying before," and it was sometimes necessary to refer to previous segments to clarify the subject of a given segment.

Twenty percent of the transcripts were rated by the second rater to assess inter-rater agreement. The second rater was the same graduate student who categorized the participants' responses according to cognitive processes, described above. Transcripts were systematically selected for the inter-rater 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. The ratings of match between the item problem-solving models and participant responses only required one iteration to arrive at satisfactory results, which are given in Table 5, both averaged across all items and at the individual item level. Inter-rater agreement did not vary considerably among items; the lowest kappa obtained for any item was 0.68 for items 3 and 6, with agreement of 86% and 87% respectively. Based on these results, inter-rater agreement was considered to be acceptable and evidence that the problem-solving models could be used consistently and non-idiosyncratically to rate participant responses.

Table 5: Inter-rater agreement on model-based coding

Item ratings	Kappa	Percent agreement
Item 1	.74	88
Item 2	.80	90
Item 3	.68	87
Item 4	.78	92
Item 5	.74	89
Item 6	.68	86
Item 7	.77	90
Item 8	.80	91
Item 9	.74	88
Verbalized item means	.75	89

CHAPTER 4

RESULTS

The following chapter first provides results on the analysis of possible method effects due to verbalization. Then descriptive statistics on relative use of different cognitive processes derived from the verbal protocols are provided. Results of a generalizability analysis to determine whether responses should be analyzed at the item level or averaged across items are presented. Then results of a series of regression analyses using convergent (construct-relevant) and divergent thinking processes and error categories as predictors are presented. Finally, individual items where minority ethnic participants scored significantly lower than white participants are selected and analyzed in depth for exploratory purposes.

Descriptive statistics on item difficulty for verbalized items and nonverbalized items are provided in Table 6. The correlation between the verbalized and the nonverbalized items was .795. Both the verbalized and nonverbalized part tests scores have high internal consistency given the small number of items, as shown in Table 7. The difference in average item difficulty for the verbalized items compared to the nonverbalized items was significant for all participants based on a dependent samples t test, t (24)= 4.102, $p < .001$. However, the verbalized items were also more difficult on average for the population that took all the items under standardized conditions in 1998. When the known population difficulty based on the 1998 test administration was used as a control variable to predict the item difficulties obtained in the present study, the addition of a dummy-coded variable for

Table 6: Descriptive statistics on sample item difficulties

	Verbalized item means (n = 9)	SD	Nonverbalized item means (n = 39)	SD	N
All participants (N = 25)	.47	.27	.62	.16	25
Majority ethnic (N= 16)	.58	.25	.67	.16	16
Minority ethnic (N = 9)	.28	.22	.52	.11	9
Males (N = 12)	.49	.31	.63	.17	12
Females (N = 13)	.45	.25	.60	.15	13

Table 7: Intercorrelations of part and total scores

	Total score	Verbalization part score	Nonverbalization part score
Total score (48 items)	0.869[5]	0.892	0.983
Verbalization part score (9 items)		0.753	0.795
Nonverbalization part score (39 items)			0.805

[5] Coefficient alpha appears on the diagonal.

verbalization condition yielded no significant increase in the amount of variance accounted for.

An analysis using participants' proportion correct for the verbalization items and nonverbalization items as repeated measures was performed with verbalization condition as a within-subjects factor and minority ethnic status as a between-subjects factor. As expected, a main effect was found for minority status F (1, 23) = 23.346, $p < .001$, partial eta squared = .270. The interaction between verbalization condition and minority status was also statistically significant F (1, 23)= 4.659, $p = .042$, partial eta squared = .168, indicating that minority participants found the verbalization items significantly more difficult compared to white participants. When the three non-native English speakers who were also from minority ethnic groups were omitted from the analysis, the interaction was still significant, F(1, 20) = 5.095, $p = .035$, partial eta squared = .203.

Based on separate chi-square tests, participants' perception of the verbalization condition as helpful, having no effect, or inhibiting to performance was independent of either sex or minority ethnic status.

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 8. These statistics are based on taking the number of segments belonging to each category in each item response for each person and dividing by the total number of segments in the 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

Table 8: Descriptive statistics on cognitive processes and error types

Cognitive processes	**Mean proportion**	**SD**
Inferring based on legal principles	.27	.07
Rehearsing cues	.24	.11
Decision-making	.14	.07
Nonsolution-oriented thinking	.18	.07
Test-strategizing	.05	.04
Drawing early conclusions	.03	.03
Inferring beyond givens and principles	.02	.02
Other (not categorized)	.03	.02
Error types	**Mean frequency**	**SD**
Errors of principles	11	6.2
Errors about given facts	1	1.2
Errors of prediction	2	2.0

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.

The decision to use a proportion was determined in light of the wide variability in length of the verbal responses among participants. The number of segments in the verbal responses ranged from eight to 53, and appeared to be based on individual participant differences in verbosity rather than item characteristics. To score based solely on frequency of utterances in, for instance, the legal principles category would have given a systematic advantage to participants who tended to articulate large quantities of ideas even if they included many idle or largely irrelevant ideas. The judgement was therefore made to rate according to relative thinking emphases (by using proportions) rather than according to straightforward frequencies of articulation types.

Elstein, Shulman, and Sprafka (1978) provide a precedent for considering the relative frequency of use of certain types of problem-solving processes. In their study on patient-management problems, they constructed an efficiency rating by calculating the percentage of cues selected from the problem that were critical findings for at least one hypothesis generated out of total cues selected. Their efficiency variable only related to efficiency in hypothesis formulation, by assigning lower efficiency scores to problem-solutions where a large number of extraneous clues were selected, resulting in a large denominator. The intercorrelations between the cognitive processes used as predictors in the analyses are provided in Table 9.

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 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.

Table 9: Intercorrelations among cognitive processes

	Cues	Early concluding	Outside inferences	Test-strategizing	Model-matching
Principles	-.44	-.16	.09	-.52	.25
Cues		-.40	.08	-.07	.20
Early concluding			-.01	.27	-.37
Outside inferences				-.22	-.38
Test-strategizing					-.35

These processes represented almost two-thirds of segments in person's total set of responses, on the average. The standard deviations of these averages indicate there was considerable variability among participants in the relative proportions of segments in the major categories.

The broad category of nonsolution-oriented thinking comprised 18% of segments for participants on average. The nature and preponderance of these segments (a typical segment classified as nonsolution-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.

The standard deviations of the remaining categories indicate the skewness of the distributions (since the proportions can not fall below zero). In other words, for some participants, the problem-solving thinking processes of using test-taking strategies such as deductive elimination and cue-seeking, making inferences outside the given facts and legal principles, and drawing conclusions before considering options, were seldom or never used.

A variance components or generalizability analysis was performed to help determine whether responses should be analyzed at the item level or averaged across items. The analysis of variance components estimates the contribution of different random effects to the variance of the dependent variable. In the words of Shavelson and Webb (1991), samples are considered random "when the size of the sample is much smaller than the size of the universe, and the sample either is drawn randomly or is considered to be exchangeable with any other sample of the same size drawn from the universe" (p. 11). Persons are typically

considered random effects in analysis of variance components because the purpose of the research is to generalize beyond the individual persons. Based on domain-sampling theory, particular test items are also considered random effects. When a large proportion of variance in the dependent variable is attributed to items, there are relatively large differences in the dependent variable based on item characteristics. Such results would tend to make it difficult to generalize about the dependent variable across items. On the other hand, relatively small contribution due to main effects of items would tend to support the similarity of items and the ability to generalize across items about the dependent variable.

In the present study, a one-facet generalizability analysis was performed on each cognitive process identified from the verbal protocols. The purpose of the generalizability analysis was to determine if the relative contributions of individual differences among participants and item differences to variance in the cognitive processes. The estimation method was the method based on minimum variance quadratic unbiased estimators (MIVQUE0). According to SAS documentation (SAS Institute, Inc., 2002), the MIVQUE0 method "produces unbiased estimates that are invariant with respect to the fixed effects of the model and that are locally best quadratic unbiased estimates given that the true ratio of each component to the residual error component is zero". Maximum likelihood-based estimation procedures were not used due to outliers and nonnormality in the dependent variable. Results are shown in Table 10. Most of the variability in all the cognitive processes was accounted for by individual differences among

Table 10: MIVQUE0 variance components estimates for cognitive processes

Dependent variable	Using legal principles				
	Sums of squares	df	Mean squares	Estimated variance components	Percent of total variance
Persons (p)	9.705	24	0.404	0.00376	23
Items (i)	10.006	8	1.251	0.00156	9
person*item, error (pi, e)	3.607	192	0.188	0.01109	68
total				0.01641	100
Dependent variable	Making outside inferences				
	Sums of squares	df	Mean squares	Estimated variance components	Percent of total variance
Persons (p)	0.665	24	0.0277	0.00012	6
Items (i)	0.547	8	0.0684	0.00003	1
person*item, error (pi, e)	0.480	192	0.0025	0.00200	93
total				0.00215	100
Dependent variable	Rehearsing facts				
	Sums of squares	df	Mean squares	Estimated variance components	Percent of total variance
Persons (p)	22.178	24	0.924	0.01028	44
Items (i)	16.367	8	2.046	0.00287	12
person*item, error (pi, e)	5.067	192	0.0264	0.01014	44
total				0.02329	100
Dependent variable	Using test-taking strategies				
	Sums of squares	df	Mean squares	Estimated variance components	Percent of total variance
Persons (p)	2.563	24	0.107	0.00085	17
Items (i)	0.652	8	0.0815	0.00000	0
person*item, error (pi, e)	1.127	192	0.00587	0.00425	83
total				0.00510	100
Dependent variable	Drawing early conclusions				
	Sums of squares	df	Mean squares	Estimated variance components	Percent of total variance
Persons (p)	1.835	24	0.0765	0.00071	24
Items (i)	1.116	8	0.140	0.00014	5
person*item, error (pi, e)	0.650	192	0.00339	0.00209	71
total				0.00294	100

persons or by the error term, which included the person-by-item interaction. Individual participants varied considerably in their use of legal principles, fact rehearsing, and tendency to draw early conclusions. With all dependent variables except rehearsal of facts, variability among test form items contributed less than 10% to overall variance. These results supported the decision to perform the remaining analyses for the study by averaging cognitive processes for each person across items, and using total scores and subscores (rather than a dichotomous outcome score for performance on each separate item) as dependent variables. Similar results were obtained in generalizability analyses on the error categories, and are presented in Table 11.

Three separate analyses were performed to assess the relative effect on performance of types of cognitive processes that were predicted to be highly construct-relevant as compared to cognitive processes that were unpredicted or considered potential sources of construct irrelevant variance. In the first analysis, the summed score on the nine verbalization items was the dependent variable. In the second analysis, the dependent variable was the summed score on the 39 nonverbalized items. In the third analysis, a composite score representing the standardized value of the equally weighted sum of the two part scores was the dependent variable. This analysis provided a more reliable dependent variable than the previous analyses and tested the predictiveness of thinking processes reported in the sample of items on performance in general.

In each linear multiple regression analysis, the construct-relevant predictor was proportional use of inferences based on legal principles; the construct-irrelevant thinking processes were

Table 11: MIVQUE0 variance components estimates for error types

Dependent variable	Errors in using legal principles				
	Sums of squares	df	Mean squares	Estimated variance components	Percent of total variance
Persons (p)	913.84	24	38.077	0.30190	15
Items (i)	1240	8	155	0.18745	9
person*item, error (pi, e)	441.76	192	2.301	1.51366	76
total				2.00301	100
Dependent variable	Errors in drawing early conclusions				
	Sums of squares	df	Mean squares	Estimated variance components	Percent of total variance
Persons (p)	87.04	24	3.627	0.02287	10
Items (i)	81.556	8	10.195	0.00843	4
person*item, error (pi, e)	50.782	192	0.264	0.19713	86
total				0.22843	100
Dependent variable	Errors in basic facts				
	Sums of squares	df	Mean squares	Estimated variance components	Percent of total variance
Persons (p)	31.84	24	1.327	0.00000	0
Items (i)	96	8	12	0.01120	5
person*item, error (pi, e)	45.76	192	0.238	0.19991	95
total				0.21111	100

proportional rehearsal of cues, 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. In all cases, the model was statistically significant and accounted for a large proportion of variance in the dependent variable. In two out of three regressions, the variable for using inferences based on legal principles was a statistically significant predictor of performance ($\alpha = .05$) on the dependent variable, with a positive coefficient. When the verbalized total or the equally weighted composite score was the dependent variable, using inferences outside facts and principles was a significant predictor with a negative coefficient. Results for the three analyses are provided in Table 12.

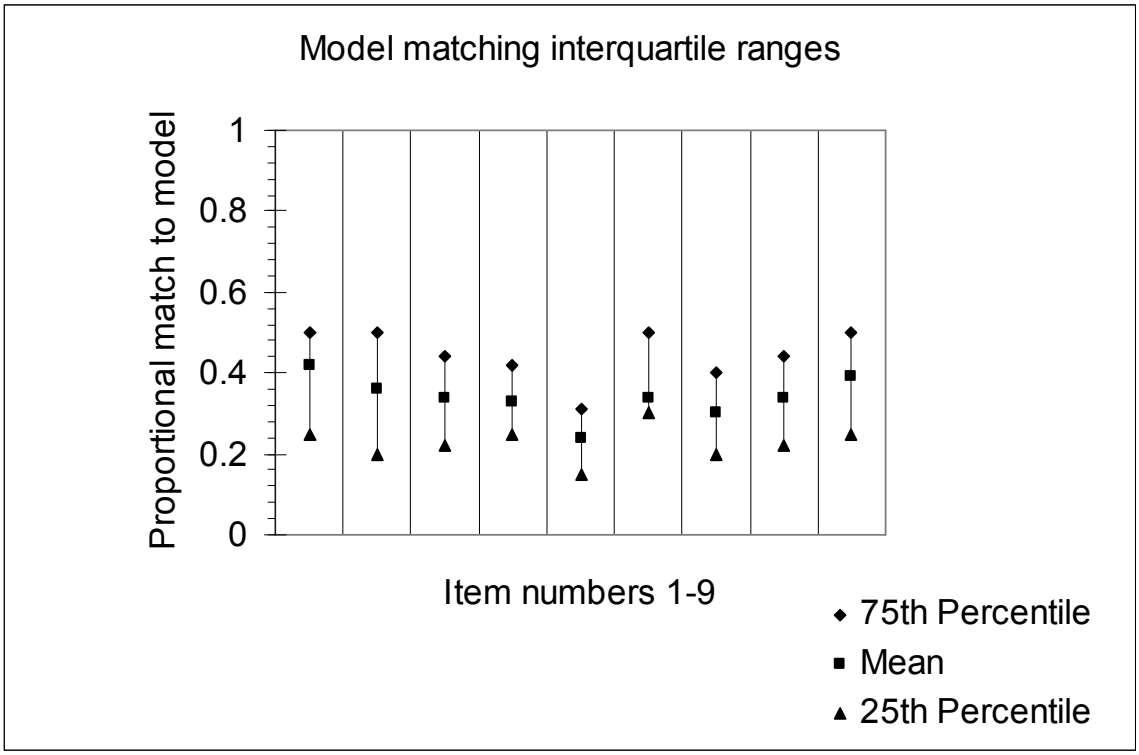
Three separate analyses were also performed to assess whether, when variance in performance due to similarity of responses to expert problem-solving models was accounted for, the use of the divergent mental processes added to prediction. Convergent processes used as predictors were operationalized as percentage of model matched by the response. Low scores on the model-matching variable represented large errors of omission or incomplete problem analyses; high scores represented relatively comprehensive analysis. The mean proportion of the model for each item matched by a participant's response and the interquartile ranges are shown in Table 13. Only rarely did participants verbalize even half of all the key points identified by the experts. This figure thus shows not only the variability in participants' responses, but their relatively low level of expertise when compared to experts.

Table 12: Summary results of regressions of cognitive processes on MBE performance measures

Dependent variable	Independent variables	B-weight	Std. error	Beta	p	Adjusted R^2(model)
Verbalized item total	Principles	15.46	8.56	0.44	0.09	0.36
	Facts	-2.27	5.29	-0.098	0.67	
	Test-taking	-17.34	14.38	-0.26	0.24	
	Early conclusions	5.94	15.71	0.074	0.71	
	Outside inferences	-50.80	22.82	-0.38	0.04	
Nonverbalized item total	Principles	72.45	19.14	0.83	0.001	0.48
	Facts	9.18	11.84	0.16	0.45	
	Test-taking	1.95	32.17	0.012	0.95	
	Early conclusions	15.03	35.14	0.075	0.67	
	Outside inferences	-87.08	51.03	-0.26	0.10	
Composite score[6]	Principles	9.50	3.16	0.67	0.007	0.46
	Facts	0.30	1.95	0.032	0.88	
	Test-taking	-3.54	5.31	-0.13	0.51	
	Early conclusions	2.56	5.80	0.079	0.66	
	Outside inferences	-18.30	8.42	-0.34	0.043	

[6] Composite score was standardized score based on equally weighted sum of standardized scores from verbalized and nonverbalized part scores.

Table 13: Means and interquartile ranges for model-matching on verbalized items



Divergent processes were operationalized as the unpredicted cognitive process of drawing early conclusions, and the construct-irrelevant processes of test strategizing and making inferences outside facts and principles. Each analysis was conducted in two stages, first by entering the average proportion of elements in the expert problem-solving model matched in participant's responses, and second by adding as a block to the regression equation the divergent processes. The hypothesis was that variability in performance would be significantly accounted for by the model-matching variable, and that construct-irrelevant and unpredicted thinking processes would not account for significant additional variance. The analyses used the verbalization item summed scores, the nonverbalization item summed scores, and the equally weighted composite scores as dependent variables.

In the regressions using the verbalization item total and the equally weighted composite score as dependent variables, the first model containing only the construct-relevant independent variable was statistically significant, and the specific variable capturing similarity to the exemplary problem-solving models was statistically significant ($\alpha = .05$). The addition of the construct-irrelevant and nonpredicted thinking processes did not significantly add to the amount of variance accounted for in either case. When the dependent variable was the nonverbalized item score, the similarity of a participant's responses to exemplary problem-solving models for the verbalized set of items did not predict performance on the nonverbalized items. Summary results for the three analyses are provided in Table 14.

Three regression analyses were also performed to assess the impact of different types of cognitive errors on test performance,

Table 14: Summary results of regressions of convergent and divergent processes on MBE performance measures

Dependent variable	Model	Independent variables	B-weight	Std. error	Beta	p	R^2 (model)	Significance of R^2 change
Verbalized item total	I	Model-matching	12.56	3.60	0.59	0.002	0.35	0.002
	II	Model-matching	11.02	4.59	0.52	0.026	0.46	0.292
		Test-taking	-20.26	13.14	-0.30	0.14		
		Early concluding	19.74	14.50	0.25	0.19		
		Outside inferences	-20.67	26.65	-0.15	0.45		
Nonverbalized item total	I	Model-matching	19.33	10.34	0.36	0.07	0.13	0.074
	II	Model-matching	10.35	13.72	0.19	0.46	0.22	0.54
		Test-taking	-57.80	39.31	-0.34	0.16		
		Early concluding	9.55	43.38	0.048	0.83		
		Outside inferences	-59.12	79.72	-0.18	0.47		
Composite score[7]	I	Model-matching	4.34	1.56	0.50	0.01	0.25	0.01
	II	Model-matching	3.24	2.03	0.38	0.13	0.35	0.41
		Test-taking	-9.27	5.81	-0.34	0.13		
		Early concluding	5.03	6.41	0.16	0.44		
		Outside inferences	-9.49	11.78	-0.17	0.43		

[7] Composite score was standardized score based on equally weighted sum of standardized scores from verbalized and nonverbalized part scores.

again using verbalization scores, nonverbalization scores, and composite scores as dependent variables. 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, 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.

The results of this analysis of error types are provided in Table 15. For all dependent variables, the model accounted for a large proportion of variance in scores. As predicted, errors of principle were significant ($\alpha = .05$) while errors of fact accounted for little variability. Errors in drawing early conclusions were also found to be statistically significant negative predictors of performance for all dependent variables.

As with large-scale MBE administration historically, a performance gap was found between white and non-white participants in the present study. Significant differences in performance were found

Table 15: Summary results of regressions of error types on MBE performance measures

Dependent variable	Independent variables	B-weight	Std. error	Beta	p	Adjusted R^2(model)
Verbalized item total	Errors in legal principles	-0.28	0.06	-0.69	<0.001	0.46
	Errors about given facts	0.24	0.34	0.11	0.48	
	Errors in early conclusions	-0.57	0.20	-0.44	0.01	
Nonverbalized item total	Errors in legal principles	-0.69	0.13	-0.69	<0.001	0.64
	Errors about given facts	1.31	0.68	0.25	0.07	
	Errors in early conclusions	-1.98	0.41	-0.61	<0.001	
Composite score[8]	Errors in legal principles	-0.12	0.022	-0.73	<0.001	0.61
	Errors about given facts	0.16	0.12	0.19	0.17	
	Errors in early conclusions	-0.29	0.069	-0.55	<0.001	

[8] Composite score was standardized score based on equally weighted sum of standardized scores from verbalized and nonverbalized part scores.

between white and nonwhite participants both in the verbalized items, t (23) = 2.93, p = .007, and in the nonverbalized items, t (23) = 2.39, p = .025. In exploratory analyses, differences in minority status predicted significant amounts of variance in total scores even when differences due to cognitive processes had been accounted for. However, when differences due to frequency of errors in using legal principles, errors of fact, and errors in drawing early conclusions had been accounted for, differences in minority status did not account for significant additional variance in total performance. These results suggest that the differences in performance between whites and nonwhites in this study were more strongly related to differences in frequency of error-making between the two groups than to any differences in their use of types of thinking processes.

For instance, whites and nonwhites did not exhibit any differences in tendency to draw early conclusions. The mean difference between nonwhites and whites on the proportional use of drawing early conclusions was 0.0096 or about 1%, and was not statistically significant. Since on average white participants' performance was considerably higher than nonwhites', this finding runs somewhat contrary to literature on problem-solving expertise, in which building hypotheses is associated with expert problem-solving. Among participants taking the selected MBE items, high-scoring participants did not tend to generate early hypotheses with greater frequency than low-scoring participants. The correlation between drawing early conclusions and total score was -0.092 and was not statistically significant.

This apparent contradiction relates to an important ambiguity in the behavior of drawing early conclusions, which made it difficult to classify the category as relevant or irrelevant to the construct of legal reasoning. On the one hand, attempting to draw conclusions prior to reviewing the multiple-choice options may be a pragmatic problem-solving tactic for those highly able to predict possible outcomes from the fact situation. If so, one would expect to find it associated with more proficient examinees. On the other hand, the behavior might also be favored by participants lacking thoroughness, who tended to gloss over important small details. Then one would expect it of lower-scoring examinees. The insignificant association between drawing early conclusions and overall performance suggests that both types of participants were drawing early conclusions.

In light of the performance gap, exploratory logistic regressions were conducted to identify items that would be particularly fruitful for further analysis. The gap between white and nonwhite performances was found to be significant ($\alpha = .10$) on five items: 1, 6, 7, 8, and 9. Chi-square analyses were also conducted on each item to see if response choice was related to minority status. In general, the items that had significant chi-square results were the same as the items with significant performance gap differences due to minority status: items 1, 5, 7, 8, and 9. Items 1, 7, and 9 will be discussed here in detail, because they were items where white and nonwhite participants showed significantly difference performance outcomes as well as distinctly different patterns of responding.

For items 1, 7, and 9, exploratory analyses conducted at the item level using the same predictors as in the previous quantitative

analyses did not explain differences in performance for minority status participants. There were no significant differences between white and nonwhite participants in the various cognitive processes studied or in the frequency of errors of different types. However, as stated above, there was a significant relationship between response choice and minority status. These item results thus illustrate that when verbal reports are quantified, the methods of categorization of the data constrain the inferences that can be made from them. Fortunately, the participants' response choices and completed uncoded transcripts are also available for study, so that if information was reported but not used in previous analyses due to the quantification method, it may still be found.

For all three items, the option most favored by nonwhite participants was option A. The response patterns for white and nonwhite participants on the three items are provided in Table 16. Approximately three times as large a proportion of nonwhites as whites chose option A for item 1, about twice as large a proportion of nonwhites as whites chose A for item 7, and about four times as large a proportion of nonwhites as whites chose option A for item 9. When white participants with overall low performance on the verbalization items (operationalized as verbalization scores lower than 50%) were compared to nonwhites at the same performance level on the verbalization items, the pattern remained similar. On item 1, 71% of low-performing nonwhites chose the first option compared to 50% of low-performing whites; on item 7, 71% of low-performing nonwhites chose the first option compared to 17% of low-performing whites; on item 9, 86% of low-performing nonwhites chose the first option compared to 33% of low-performing whites. In all these

Table 16: Response patterns for white and nonwhite participants on items 1 and 9

	Option	Item 1	Item 7	Item 9
Whites (n=16)	A	3	6	3
	B	1	0	3
	C	11[9]	3	9*
	D	1	7*	1
Nonwhites (n=9)	A	5	6	7
	B	2	1	0
	C	2*	2	2*
	D	0	0*	0

[9] Indicates correct response

cases, the option that was incorrect but disproportionately preferred by nonwhites was the first option in the response list. This may indicate an unconscious preference on the part of the nonwhite cases, the option that was incorrect but disproportionately preferred by nonwhites was the first option in the response list. This may indicate an unconscious preference on the part of the nonwhite participants in the study for the first plausible option. A preference based on primacy would not have been captured in the verbal protocols; it would be exactly the sort of unconscious psychological tendency not consciously available to subjects.

Qualitative exploration of the verbal protocols did not provide conclusive information as to whether nonwhite participants unconsciously had a stronger preference for the first option in the response list than white participants. None of the nonwhite participants articulated a primacy preference, although one stated, "I'm tired, I'm just going to pick A." Also, one nonwhite participant, finding A to be a satisfactory option for item 1 because it established negligence *per se,* did not go on to read the other options or consider the problem of causation.

In the cases of items 1 and 9, primacy is confounded with partial correctness. Both items 1 and 9 were in the content area of torts and involved a wrongful death action. The principal elements that must be established by a plaintiff in a wrongful death action are duty, breach of duty (which together constitute negligence), causation, and harm or damages. In each item, option A presented an argument based on the establishment of negligence only, while the correct option articulated the need for "but for" causation. "But for" causation is the principle

that “but for” the negligence, the harm would not have occurred. Thus in these two items, the choice of option A may indicate a tendency for the nonwhite participants in the study to foreclose on a simplified response that satisfies part of the conditions of the legal issue in question. An example of such a tendency is found in the verbalization of the nonwhite participant who stated, “I don’t like that C is a but for clause and they always seem to complicate the issue so I’m hoping that one of the other ones is clear.” The participant eventually selected option A.

If the same mechanism relating to the response preference among nonwhites is at work in all three of these items, item 7 would seem to argue in favor of a position or primacy effect as opposed to a preference for a partial solution. Item 7 concerns a takings issue in the constitutional law area. The fact situation presents a scenario in which a city government requires a commercial property owner to build a daycare center as a condition for issuing a permit for a proposed building expansion. The requirement is authorized by local zoning regulations, which are challenged by the property owner. In court, the city provides no evidence of need for daycare facilities. This item and the expert problem-solving model were provided in Chapter 3 of this report. The first option for Item 7 does not provide a partial solution to the problem because it relates to the rational basis test for constitutionality, which is an incorrect standard of review for a takings issue. However, examination of the verbal reports of the low-scoring shows that not a single low-scoring participant verbally identified the problem as relating to an issue of takings. In the absence of declarative knowledge about the issue, the rough

proportionality test for constitutionality that was actually relevant to the problem appeared strange and unfamiliar. In the words of one participant, "…that rough proportionality, I've never heard the word rough as a term of art in law…" and in the words of another, "that doesn't sound like the right terminology at all." Most participants eliminated the correct answer early, and also eliminated option B, which contained a factual error about the legal rules for the rational basis test in that it placed the burden of proof on the wrong party. For reasons that remain unclear, the low-scoring whites were more tempted by option C which was correct in its conclusion about unconstitutionality of the takings, but incorrect in level of review, while low-scoring nonwhites were more tempted by option A which was incorrect in its conclusion but required a lower standard of review than option C.

In most cases, because of the high plausibility of the item distractors, participants were able to provide a principle-based reason for their choice, even though the use of the principle was incorrect or only partially correct for that fact situation. Thus examination of the verbal protocols does not reveal if the differences in response patterns for these items indicate greater difficulty for nonwhite participants in grasping specific key principles (e.g., "but for causation"), or a greater unconscious primacy effect hidden beneath rationalizations.

CHAPTER 5

DISCUSSION

This dissertation presents literature relevant to defining the construct of legal reasoning and identifying likely cognitive processes involved in problem-solving on restricted MBE-type examination items. It provides evidence about the suitability of gathering information about the examinee mental processes through analysis of verbal reports. The method as applied in the current study is described in detail.

Results from analyses of method effects due to verbalization, and regression analyses using convergent and divergent thinking processes and error categories as predictors of performance are presented. Individual items where minority ethnic participants scored significantly lower than white participants are selected and analyzed in depth for exploratory purposes.

The following conclusions are drawn based on the study. First, the method of verbal protocol analysis seems to be well suited to this type of task and this type of examinee. The selected MBE items are of moderate but not overwhelming complexity, presenting frequent cues to call up associations of declarative knowledge from long-term memory into short-term memory where they are available for verbalization. The majority of participants in the study were highly verbal, and seemed comfortable talking aloud as they worked through the selected items.

The study shows that it is possible to develop categorical systems for classifying mental processes reported in verbal protocols both prior to coding and during pilot stages of the coding process itself. These systems can be used with high reliability if the raters are well trained in the content relevant to the verbalization stimuli.

In the case of this study, the raters needed to develop considerable expertise in the legal principles potentially at issue relevant to the nine items used as stimuli.

The study provides insight into the mental processes at work in individuals as they respond to items such as those used on the MBE. Participants in the study engaged in a variety of thinking activities, dominated by thinking about legal principles. The relatively large proportion of verbalizations devoted to rehearsing cues from the fact situations and answer options suggests that the fact situations in MBE items are complex and highly detailed. Participants also engaged in a considerable amount of thinking not clearly productive of problem solutions, including completely irrelevant mental wandering and mental "floundering," apparently during attempts to retrieve information from memory. Test-taking strategies and other types of thinking were observed less frequently.

The results of the regression analyses performed in this study provide evidence in support of 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. Three types of analyses converge on strikingly similar results: when proportional use of the different types of cognitive processes is used to form predictors, the application of legal principles has by far the strongest positive relationship to performance. High performance on the selected items is associated with greater proportional use of construct-relevant thinking processes (making inferences about legal principles) and has no significant relationship to potential sources of construct-irrelevant variance such as reading comprehension and using

test-taking strategies. In general, using the construct-irrelevant strategy of making inferences outside facts and principles is associated with low performance on the selected items. Making inferences outside facts and principles represents a nonlegalistic tendency to reason from beliefs or intuition rather than evidence, and is a tactic associated with individuals of lower proficiency on the selected items. Lastly, drawing early conclusions has no statistically significant relationship to performance. Whether this final finding from the regression analyses of cognitive processes supports or detracts from the interpretation of test scores as measures of legal reasoning will be discussed below.

The use of an alternative coding approach that developed a quantitative variable for the similarity of participants' responses to expert problem-solving models and conveyed to some degree errors of omission in participant thinking, provides confirmation of the stability of these results. The value of reasoning using thoughts highly relevant to problem solutions and the relative lack of value for divergent types of thinking is similarly supported by the analyses using expert problem-solving models.

Analyses of the contribution of different types of errors to performance support the importance of correct construct-relevant knowledge, and the relative irrelevance of construct-irrelevant variability in reading comprehension. Errors in early conclusions negatively impact performance on the selected MBE items.

Evidence was found that differences in performance between white and nonwhite examinees are more associated with differences in frequency of errors rather than differences in test-taking strategies

or mental processes. However, in this case, the method of the study presents questions that require further investigation. The interpretability of the responses of the nonwhite participants was compromised by an apparent method effect, in that nonwhite participants found the verbalization items disproportionately more difficult than did the white participants. The cause for this disproportionate difficulty is unknown, except that it does not appear to be related to differences in English language acquisition.

A plausible hypothesis to explain the effect can be found in the research on stereotype threat originated by Steele and Aronson (1995). Stereotype threat occurs when participants are aware of being at risk of confirming a negative stereotype about their own population group. For instance, when informed that a test was intended to be diagnostic of verbal ability, a random selection of African Americans participants underperformed in relation to Whites; however, when the identical test was administered without the information that it was diagnostic of abilities associated with racial differences, African Americans did not underperform compared to Whites.

It is likely that nonwhite participants in the present study were aware of the minority status MBE performance gap. Although instructions for the selected item verbalizations did not suggest that performance in the study would be predictive of future MBE performance, many participants were interested in the study precisely because they viewed it as a way to estimate their own future performance. According to theory, stereotype threat causes "an inefficiency of processing much like that caused by other evaluative pressures. Stereotype-threatened participants spent more time doing fewer items more inaccurately"

(Steele and Aronson, 1995, p. 809). This type of debilitation has also been shown as a reaction to the presence of an audience. It may be that stereotype threat and pressure under observation combined to create the verbalization condition-by-minority status interaction observed here. It is perhaps worth noting that both the researcher in the present study and the assistant interviewer are white, and that interviewer effects have been observed in other verbalization studies.

There is an alternative explanation for the disproportionate difficulty of the verbalization items for minority participants. There were three important differences between the verbalization item set and the nonverbalization item set: the verbalization items were answered aloud, potentially giving rise to the increased anxiety and stereotype threat described above; the verbalization items were on average more difficult; the verbalization items were on average more discriminating. If one imagines analyzing these two sets of items using item response theory, the more discriminating items would have steeper slopes around the inflection point of their item characteristic curves than the nonverbalization items. This means that the more discriminating set of items would have relatively large gaps in difficulty (or probability of correct responses, change along the y-axis) associated with small differences in participant ability (change along the x-axis). In comparison, the less discriminating, nonverbalization items would have smaller differences in difficulty associated with differences in ability. Thus because the minority participants were on average slightly lower in ability than the non-minority participants, the measurement of this difference is greatly heightened when items are more discriminating. The apparently disproportionately greater

difficulty of the verbalization items for the nonwhite study participants may be an artifact of the verbalization items's greater ability to discriminate between levels of ability.

Analyses of individual item responses where nonwhite participants had significantly more difficulty than white participants are inconclusive, suggesting either a tendency in the nonwhite participants to settle on partially correct responses, or to select plausible responses that appeared early and rationalize about them. It would be desirable for further studies to test these rival hypotheses about the cause of the observed unusual response patterns in nonwhite participants through experimental manipulation with items designed to stimulate the predicted behaviors. For the present, all that is clear is that if nonwhite and white participants were responding differently to the test items, the present method of thinking-aloud did not reveal differences in their reasoning.

Overall, the study finds that primarily construct-relevant thinking positively predicts performance on the selected MBE items, and construct-irrelevant thinking either detracts from or is relatively unimportant to performance on the selected items. The process of drawing early conclusions was not categorized as clearly construct relevant or irrelevant, and the results about its relationship to performance are ambiguous: in general, it has no statistically significant relationship to performance on the selected items, but making incorrect judgments through the use of this type of thinking has a strong negative impact.

The fact that the tendency to draw early conclusions is not associated with high performance overall runs contrary to some research

on expert problem-solving, in which experts are found to engage in building hypotheses about problems early in the solution process. Three explanations for this result are suggested. First, the expert problem-solving literature that has found early hypothesis-building in experts but not novices is based primarily on open-ended tasks. While novices may delay building hypotheses on open-ended tasks, they may to behave differently on multiple-choice items, especially under time constraints. On timed multiple-choice tests such as the MBE, novices and experts alike may be motivated to attempt to draw conclusions early. This study found that participants at both ends of the performance continuum drew early conclusions. Some early conclusions were premature ones based on superficial analyses, while others were the results of sophisticated analyses and relatively expert insight.

Second, few participants in this sample may have been near the level of expertise where early hypothesis-building would be a very useful tactic. A relatively low level of expertise would be expected even among high scorers on selected items from a minimal competency licensure exam. The kind of deep structural insight into problems necessary for building accurate early hypotheses may have been largely out of the reach of the study participants.

Third, the MBE tasks themselves present pitfalls for those who rush to conclusions early. Some of the literature on legal reasoning suggests that developing hypotheses before all facts are absorbed is associated with fallacious reasoning. MBE items seem actively to discourage such fallacious tendencies in favor of thoroughness, by sometimes introducing new, conditional facts in response options. Drawing early conclusions will be a problem in MBE items when new facts

are presented in the response options that alter the interpretation of the fact situation, and that may be overlooked if a premature decision has been made.

This characteristic of MBE items apparently rewards thoroughness and attention to detail. The construct of legal reasoning is a complex one which experts agree includes not only the dogmatic ability to apply legal principles meticulously, but also the ability to reinterpret legal rules and principles on a case-by-case basis in the pragmatic light of public policy needs. Given the two strands of thinking in the legal literature of dogmatism and pragmatism, the MBE items appear to emphasize the more dogmatic approach. However, time constraints may impel test-takers to draw conclusions early despite possible pitfalls in the answer options. Moreover, some would consider that when examinees analyze a fact situation and draw conclusions independent of the options, they use a more sophisticated kind of analysis than when they analyze each option for its applicability to the facts. Presenting new facts or extenuating circumstances in the options discourages more predictive and synthetic analysis in favor of dogmatic application of rules.

The question of how well the application of legal principles as measured by the MBE represents the overall construct of legal reasoning is beyond the scope of the present study. However, the tension between the dogmatic and pragmatic strands in the overall construct is evident when the process of drawing early conclusions is examined. The results of the study suggest that MBE scores should be interpreted as more purely dogmatic applications of legal rules, because the MBE item format discourages the pragmatic and sophisticated attempt to analyze

problems and draw conclusions early by allowing essential qualifiers to be inserted in the multiple-choice options.

The study has a number of limitations, many of which could be addressed in future research. The study method may have produced reduced veridicality in the reports by the lack of speededness in the verbalized items. It may be, for instance, that examinees taking the timed MBE draw early conclusions more frequently than the participants in the present study did, knowing that time was not an issue. It may be possible to estimate the factor by which verbalizing responses increases response time on items of this type, and impose a modified time limit for verbalization items to provide greater authenticity in a future study.

Also, the large proportional use of non-solution oriented thinking in the present study suggests mental processing not captured by the verbalization method, at least not as it was applied in this study. To find out if the type of thinking underlying these inconclusive statements might be available to participants in retrospect, a study that combined concurrent verbalization with guided retrospection could be conducted, as recommended in Leighton (2005).

Finally, the study is limited in its generalizability due to participant and item sampling. First, because of the low sample size the regression analyses are unstable estimators of population parameters, and the results may be highly influenced by sampling error. Also, the study was based on items from only three of the six MBE content areas, and only highly-discriminating items were selected for the study. A study that included more low-discriminating items might

find performance on those items to be more affected by test-taking strategies, for instance.

It is important to state that the absence of evidence is not the evidence of absence, and that some of the construct-irrelevant predictors had low incidence and variability in the study, which impacted the relative strength of potential correlations with outcomes. With larger sample size, other predictors would very likely have reached statistically significance, and the direction of their effect could be gauged. Given the laborious nature of think-aloud studies, a sample size large enough to obtain stable estimates of rarely-used strategies might be difficult to achieve. The method of verbal protocol analysis appears to be useful at least for gauging relative effects if not absolute parameter estimates.

Overall, this dissertation represents a serious attempt to gather evidence about mental processes of highly motivated persons attempting to solve problems similar to those used on a nationally administered high stakes professional licensure exam. Bloom and Broder (1950) wrote more than half a century ago about the importance of the research on students' mental processes as follows:

> Much difficulty can be anticipated in securing evidence about the processes, and perhaps the nature of the human organism is such as to prevent the securing of ay clear-cut and objective evidence on these processes. In any case, attention to the processes of thought must mean the development of new techniques for psychological research. It may also require a change from large-scale testing and mass studies to those which involve small numbers of subjects studied by rather intensive techniques. The question of whether such research would be fruitful, or even possible, can be answered only after many serious attempts have been made. The challenging nature of the problem and the tremendous possibilities which would arise from a successful attack and solution should serve to channel much of our research effort to this field.

This dissertation represents one response to Bloom and Broder's challenge. It is hoped that the results about the relative impact of different types of cognitive processes on selected items from the Multistate Bar Exam are persuasive, and that this study has shown the usefulness of asking examinees to think aloud as a way of validating the interpretation of test scores in terms of constructs.

REFERENCES

Aanstoos, C. M. (1983). The think aloud method in descriptive research. *Journal of Phenomenological Psychology,* 14 (2), 243-266.

Aldisert, R. J. (1997). *Logic for lawyers: A guide to clear legal thinking* (3rd ed.). South Bend, IN: National Institute for Trial Advocacy.

Bloom, B. S., & Broder, L. J. (1950). *Problem-solving processes of college students.* Chicago: University of Chicago Press.

Bond, L. (1990). Understanding the black/white student gap on measures of quantitative reasoning. In P. Serafica, A.I. Schwebel, P. D. Isaaac., R. K. Russell, & L. B. Meyers (Eds.), *Mental health of ethnic minorities* (pp. 89-107). New York: Prager.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (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.

Case, S. M., & Ripkey, D. R. (2005, April). *Enhancing construct representativeness of licensure examinations by using multiple formats*. Paper presented at annual meeting of the National Council on Measurement in Education, Montreal, QC.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences,* 6 (3), 271-315.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5,* 121-152.

Chung, G. K. W. K., de Vries, L. F., Cheak, A. M., Stevens, R. H., & Bewley, W. L. (2002). Cognitive process validation of an online problem solving assessment. *Computers in Human Behavior,* 18, 669-684.

Clement, J. (1989). Learning via model construction and criticism: Protocol evidence on sources of creativity in science. In J. Glover, R. Ronning, R., and C. Reynolds (Eds.), *Handbook of creativity: Assessment, theory and research* (pp. 341-381). New York: Plenum.

Complete guide to law schools (2005). New York: U.S. News and World Report L. P.. Retrieved from worldwide web on April 23, 2005, at www.usnews.com.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2rd ed., pp. 443-507). Washington, DC: American Council on Education.

De Tocqueville, A. (1975). *Democracy in America* (J. P. Mayer, Trans.). Garden City, NY: Doubleday. (Original work published in 1835).

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving: An analysis of clinical reasoning.* Cambridge, MA: Harvard University Press.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: MIT Press.

Ericsson, K. A. (1987). Theoretical implications from protocol analysis on testing and measurement. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), *The influence of cognitive psychology on testing* (pp. 191-228). Hillsdale, NJ: Erlbaum.

Friedman, L. M. (1985). *A history of American law.* New York: Simon & Schuster.

Hamilton, L. S. (1994, April). *Validating hands-on science assessments through an investigation of response processes.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview Procedures for Validating Science Assessments. *Applied Measurement in Education*, 10 (2), 181-201.

Harte, J. M., Westenberg, M. R. M., & van Someren, M. (1994). Process models of decision making. *Acta Psychologica,* 87, 95-120.

Heller, J. I., Sheingold, K., & Myford, C. M. (1998). Reasoning about evidence in portfolios: Cognitive foundations for valid and reliable assessment. *Educational Assessment,* 5 (1), 5-40.

Kane, M. T. (2005). The role of licensure tests. *The Bar Examiner, 74*, 1, pp. 27-38.

Katz, I. R., Bennett, R. E., & Berger, A. E. (2000). Effects of response format on difficulty of SAT-Mathematics items: It's not the strategy. *Journal of Educational Measurement, 3*7 (1), 39-57.

Klein, S. P. (1993). *Summary of research on the multistate bar examination.* Chicago: National Conference of Bar Examiners.

Leighton, J. P. (2005). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational measurement: Issues and practice,* 23 (4), 6-15.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). Washington, DC: American Council on Education & National Council on Measurement in Education.

Newell, A. & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review,* 84 (3), 231-259.

Norris, S. P. (1990). Effect of eliciting verbal reports on thinking on critical thinking test performance. *Journal of Educational Measurement,* 27 (1), 41-58.

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher,* 18 (1), 16-25.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading.* Hillsdale, NJ: Erlbaum.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition,* 17 (6), 759-769.

SAS Institute Inc. (2002). SAS OnlineDoc® 9. Cary, NC: Sas Institute Inc.

Schoenfeld, A. H., & Herrmann, D. J. (1982), Problem perception and knowledge structure in expert and novice mathematical problem solvers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8* (5), 484-494.

Shavelson, R. J., & Webb, N. W. (1991). *Generalizability theory: A primer*. Thousand Oaks, CA: Sage.

Shulman, L. S., & Elstein, A. S. (1975). Studies of problem-solving, judgment, and decision making: Implications for educational research. *Review of Research in Education*, 3, 3-42.

Smith, A. G. (1977). *Cognitive styles in law schools.* Austin: University of Texas Press.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69* (5), 797-811.

Stevens, R. (1983). *Law school: Legal education in America from the 1850s to the 1980s.* Chapel Hill: University of North Carolina Press.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes.* London: Academic Press.

Vandevelde, K. J. (1996). *Thinking like a lawyer: An introduction to legal reasoning.* Boulder, CO: Westview Press.

Wightman, L. F. (1998). *LSAT national longitudinal bar passage study.* Newtown, PA: Law School Admission Council.

Yepes-Baraya, M. (1996, April). *A cognitive study based on the National Assessment of Educational Progress (NAEP) science assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.