Evidence Evaluation in Complex Decision Making

Nancy Pennington
Graduate School of Business
University of Chicago

Reid Hastie
Northwestern University

This study investigates the role of representation of evidence in a juror's decision process. A model is presented that includes an initial stage of processing in which cognitive representations of the evidence in the form of stories are produced. This is followed by a computation of the decision by evaluating the goodness-of-fit of the evidence representation (story) to the verdict categories. Subjects, drawn from jury pools, made individual decisions on the verdicts for a filmed murder trial. Extensive interviews provided the data for analysis of their cognitive representations of the evidence in the case, the verdict categories presented in the trial judge's instructions, and the procedures they were to follow according to law to reach a verdict. We found, as hypothesized, that the trial evidence was represented in a story form. Furthermore, differences among jurors in cognitive representations of evidence were correlated with their verdicts, although other aspects of the decision process (verdict category representations, application of the standard of proof procedural instruction) were not. We conclude that adequate theories of decision making must emphasize cognitive aspects of performance, such as the representation of evidence.

In many important decisions, the decision maker must review and evaluate a large base of interdependent items of evidence before selecting an action or attitude from the available alternatives. Comprehension and evaluation of the evidence may dominate these decisions in terms of effort, time spent, and influence on the outcome relative to other aspects of the decision process. The potential significance of evidence evaluation is especially striking in legal decisions owing to legal conventions of instructing the juror to suspend judgment, following elaborate rituals of evidence presentation, and providing decision alternatives (verdicts) only at the end of the trial.

We have argued previously (Pennington & Hastie, 1981a) that an adequate model of decision making for complex decision tasks such as a juror's decision must include descriptions of the cognitive representations of evidence. The proposal stands in contrast to the emphasis of traditional decision models on information combination and computation rather than on representation. The theme of the present research is that decision making is guided in its initial stages by attempts to understand the meaning of the evidence in order to develop an interpretation or explanation for it. Subsequent phases produce action on the basis of the interpretation. The purpose of this article is to develop a specific model for evidence representation applied to an important decision task, a juror's predeliberation verdict decision.

There are four categories of traditional models that attempt to account for judgment in juror or jurorlike decision tasks (Pennington & Hastie, 1981a). (a) Information integration models (Kaplan, 1975; Kaplan & Kemmerick, 1974; Ostrom, Werner, & Saks, 1978) describe the juror's decision process as an averaging calculation integrating information on a single dimension of culpability, across an initial opinion and information items pertaining to guilt. Each evidence item is evaluated independently for implications of guilt and, at the conclusion of the trial, all evaluated items are weighted and averaged to determine a final judgment of probability of guilt. (b) Bayesian models (Marshall & Wise, 1975; Schum, 1975, 1977; Schum & DuCharme, 1971; Schum & Martin, 1982; Schum & Pfeiffer, 1973) describe the decision process as a sequence of multiplicative products of a prior opinion and the diagnosticity of each evidence item. Most applications of the Bayesian model treat evidence items as independent sources of information, but recent work by Schum and his colleagues has incorporated conditional dependencies between the implications of the items. (c) Poisson process stochastic models (Kerr, 1978a, 1978b; Thomas & Hogue, 1976) describe the decision as a constant accumulation of the weight of evidence until a critical event occurs that fixes the weight at a final value. At this time, the apparent weight of the evidence is compared to a decision criterion and the decision is made. (d) Sequential weighting models (Anderson, 1959; Einhorn & Hogarth, 1985; Lopes, 1982; Weld & Danzig, 1940; Weld & Roff, 1938) describe the process as a series of opinion revisions, where each revision is the weighted average of the previous judgment and the value of the current evidence item.

In our review, we concluded that although these models have contributed to our understanding of the juror decision process,
there are many empirical phenomena associated with realistically complex legal decisions that lie beyond these traditional approaches. For example, our experiments, interviews, observations, and intuitions about naturally occurring juror decisions suggest that the decisions do not show a continuous updating pattern across evidence presentation; that interdependencies among evidence items are ubiquitous; that inferences about missing or unstated intermediate conclusions are frequent and determinative; that there is considerable uncertainty, but that reasoning under uncertainty does not follow nor is it expressed in terms consistent with conventional probability theory; and that the penultimate stage of the decision process involves reasoning about multiattribute verdict categories rather than a unitary culpability dimension. Our efforts to account for these phenomena as well as the input–output relation between the evidence and the verdict have led us to an alternate account of juror decision making that we call the Story Model.

A Story Model of Evidence Evaluation

The Story Model is based on the hypothesis that jurors impose a narrative story organization on trial information, in which causal and intentional relations between events are central (Bennett & Feldman, 1981; Hutchins, 1980; Pennington, 1979, 1981; Pennington & Hastie, 1980, 1981a, 1981b, 1985). We propose that the story organization facilitates evidence comprehension and enables jurors to reach a predeliberation verdict decision. In overview, the Story Model includes three components: Evidence evaluation through story construction, decision alternative representation (verdict category establishment for the juror task), and story classification (selecting the verdict category that best fits the story based on the evidence).

Story Construction

This stage occurs during and after the presentation of trial evidence. Meaning is assigned to trial evidence through the incorporation of that evidence into one or more plausible accounts or stories describing "what happened" during events testified to at the trial. General knowledge about the structure of human purposive action sequences, characterized as an episode schema, serves to organize events according to the causal and intentional relations among them as constructed by the juror.

Our hypothesis concerning the organization of evidence is based on findings that narrative discourse is comprehended through a basic conceptual schema that describes the structure of most human action sequences. Although details differ among researchers, almost all accounts suggest that an "episode schema" describes a general knowledge structure representing a standard event sequence for events in both the real world and in stories. The set of relations expressed in the episode schema (see Figure 1) includes events that initiate a main character’s psychological states and goals that provide reasons for the character’s subsequent actions resulting in an outcome or consequence. Accompanying physical states may enter in two ways: either as the character’s current state at the time of the initiating events that enables subsequent actions, or as the character’s state resulting from the initiating events and contributing to the initiation of mental states or goals. The direct links shown between components (see Figure 1) in the episode schema may not all be utilized in any given interpretation or retelling of an action sequence.

The components of this schema have been labeled in various ways by other theorists (e.g., Mandler, 1980; Rumelhart, 1975, 1977; Stein & Glenn, 1979; Thorndyke, 1977; Wilks, 1977). For example, Rumelhart (1977) considers an episode schema to contain a cause schema (initiating events), a try schema (with goal, action, and consequence), and a final outcome schema (consequence). Similarly Stein and Glenn’s (1979) story grammar posits a structure in which initiating events lead to an internal response leading to a sequence of attempts that have a consequence. Mandler’s episode schema has a beginning causing a development causing an ending (Mandler, 1980; Mandler & Johnson, 1977). Theories of natural language understanding applied to narrative comprehension often implicitly advocate such sequences even though a formal grammar may not be proposed (Schank & Abelson, 1977; Trabasso, Secco, & van den Broek, 1982).

It is important to keep in mind that the knowledge structure for a story can be described at several levels of detail. For example, Schank and Abelson (1977) describe a hierarchical system with abstract themes and plans at the highest level (summarizing an entire story) and detailed conceptual dependency structures at the lowest level (corresponding to a single proposition or action). Other theoreticians have emphasized the recursive power of knowledge structures to "embed" within one another,
so that we would expect episodes to serve as elements of higher order episode structures (e.g., Rumelhart, 1977).

Our model of juror’s story construction utilizes these themes from the discourse comprehension literature in the following way: (a) The juror’s story can be represented as a hierarchy of embedded episodes; (b) at the highest level, organization can be characterized in terms of the general episode schema representing the most important features of “what happened” and bearing a direct relation to verdict alternatives; (c) components of the highest level episode may be elaborated in terms of more detailed event sequences in which causal and intentional relations among subordinate story events are represented.

It may seem inevitable that the evidence should have a story structure, a point in favor of the intuitive appeal of this theory. However, it should be remembered that the evidence is presented at trial in a disconnected question and answer format. Witnesses are typically not allowed to speculate on necessary connecting events such as why certain actions were carried out, what emotional reaction a person had to a certain event, and so forth. Different witnesses testify to different pieces of the chain of events, usually not in temporal or causal order. Even within a given witness, questions usually do not elicit information in narrative order. (A short version of our stimulus trial, shown in the Appendix, has been “narratized” within witness in order to condense 3 hours of testimony into comprehensible summary form.) Furthermore, courtroom observations and our experience with trial attorneys suggest that storytelling is not the typical summary style for an attorney’s closing argument to the jury. It is just as likely that summary remarks will be organized as a witness-by-witness review or by evidentiary issue in the form of legal argument.

Verdict Category Establishment

This stage includes the comprehension and learning of the decision alternatives. Most of the information for this processing stage is given to jurors at the end of the trial in the judge’s instructions on the law. In the juror’s decision task for criminal cases, this stage involves the representation of each verdict alternative as a category with certain defining features and a decision rule specifying their appropriate combination.

We hypothesize that the conceptual unit is a category defined by a list of criterial features referring to the identity, mental state, circumstances, and actions linked conjunctively or disjunctively to the verdict alternative (Kaplan, 1978). For example, a defendant is guilty of second-degree murder (see Table 1) if the defendant is the person who killed the victim (identity), if the defendant attacked the victim with a deadly weapon (actions) without sufficient provocation (circumstances), and if the defendant attacked the victim with the intention of inflicting an injury that could result in death (mental state). The complete representation of the verdict categories is a difficult task as illustrated in the summary (see Table 1) of the four verdict definitions for the murder trial used as a stimulus in the present research.

A similar representation has been suggested by Svenson (1979) for other categorical judgment tasks, although verdict choices have not been conceptualized in this manner in previous research on the juror’s decision task (Pennington & Hastie, 1981a). Rather, a continuous single dimension of “likelihood of guilt” or a dichotomous guilty/not guilty choice has been used. The verdict choice may reduce to a single dimension when there is only one issue in the trial, for example in a case where there is dispute over the identity of the person who committed undisputed actions with undisputed intentions in undisputed circumstances (“who done it”). However, almost all criminal cases involve more than two verdict alternatives each defined by several elements, and the judge specifically instructs the jurors that the prosecution must prove each element of the charge against a defendant beyond a reasonable doubt.

Story Classification

The story classification stage specifies that the juror’s judgment takes the form of a classification process in which the best match between story features and verdict category features is

<table>
<thead>
<tr>
<th>Verdict category</th>
<th>Identity</th>
<th>Mental state</th>
<th>Circumstances</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>First-degree murder</td>
<td>(A) Right person (B) Purpose formed</td>
<td>(A) Intent to kill (B) Interval between resolution and killing</td>
<td>(A) Insufficient provocation (B) Killing in pursuance of resolution</td>
<td>(A) Unlawful killing = Did not exhaust escape, attacker, excessive force (B) Killing</td>
</tr>
<tr>
<td>Second-degree murder</td>
<td>(A) Right person (C) Resolution to kill (B) Deliberate, cruel act</td>
<td>(A) Intent to inflict injury likely to result in death (A) Heat of sudden passion (B) Diminished capacity</td>
<td>(A) Insufficient provocation</td>
<td>(A) Unlawful killing = Did not exhaust escape, attacker, excessive force (B) Used deadly weapon</td>
</tr>
<tr>
<td>Manslaughter</td>
<td>(A) Right person</td>
<td>(A) Great provocation (B) Threat to life not immediate</td>
<td>(A) Insufficient provocation (B) Killing in pursuance of resolution</td>
<td>(A) Unlawful killing = Did not exhaust escape, attacker, excessive force</td>
</tr>
<tr>
<td>Self-defense: not guilty</td>
<td>(A) Right person (B) In defense</td>
<td>(A) Fear of life or great bodily harm (B) Unable to escape</td>
<td>(A) Under immediate attack</td>
<td>(A) Killing (B) Exhaust escape (C) Reasonable force</td>
</tr>
</tbody>
</table>
Table 2
Correspondences Between Episode Schema Components and Verdict Attributes

<table>
<thead>
<tr>
<th>Verdict attributes</th>
<th>Episode schema component</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Actions</td>
<td>Actions</td>
</tr>
<tr>
<td>2. Mental State</td>
<td>Psychological states</td>
</tr>
<tr>
<td>3. Circumstance</td>
<td>Initiating events</td>
</tr>
<tr>
<td>4. Identity</td>
<td>Implicit in actions</td>
</tr>
</tbody>
</table>

determined. The classification process is aided by relatively direct relations between attributes of a verdict category (crime elements) and components of the episode schema as shown in Table 2. It is probably not an accident that the main attributes of the decision categories suggested by legal experts (Kaplan, 1978)—identity, mental state, circumstances, and actions—correspond closely to the central features of human action sequences—initiating events, goals, actions, and accompanying states.

Because verdict categories are not well-learned concepts, the classification of a story into an appropriate verdict category is likely to be a deliberate process. Suppose that a person believes that the defendant in our stimulus trial argued with the victim (initiating event), became very angry and decided to kill him (psychological state and intention), obtained a weapon, found the victim, got in a fight and stabbed the victim (actions), resulting in the victim's death (consequence). The juror must decide which verdict category this story exemplifies by checking the mental state required by a verdict category (e.g., intent to kill for first degree murder) against the psychological states and goals of the highest level episode of the believed story (intent to kill). There is a clear match in this example (see Table 2). However, in matching verdict-required circumstances (insufficient provocation) against the story-initiating events and states (argument), the juror has to judge whether an argument is an example of insufficient provocation. This matching process continues for all verdict attributes, for each verdict category. The best-fitting verdict category is retained as a tentative decision.

The story classification stage also involves the application of the judge's procedural instructions on the presumption of innocence and the standard of proof. Although the instructions indicate that the standard of proof (beyond reasonable doubt) should be applied separately to each element of each verdict considered, the juror's evaluation is hypothesized to be global. Evaluations of plausibility, on the basis of story completeness and coherence, are preliminary. If an incriminating story is deemed plausible, then a general evaluation of goodness-of-fit of story to verdict category determines the final verdict decision.

The basic claim of the Story Model is that story construction enables critical interpretive processing and organization of the evidence such that evidence can be meaningfully evaluated against multiple judgment dimensions. The Story Model provides a psychological account for the eventual assignment of relevance to presented and inferred information. In addition, precise claims for the representational form of the evidence provide a starting point to examine the role of representational structure in assessing uncertainty in the evidence evaluation process.

Research Approach

The Story Model presented here is a hypothesis that is being examined, modified, and elaborated through a program of research. Not all features of the model can be tested simultaneously, and the initial empirical emphasis is on description of representational structures rather than on testing a detailed processing account, although the nature of proposed representations will imply certain processing assumptions. This emphasis is suggested by Rumelhart and Abrahamson (1973), who proposed that the first theoretical problem in studying a particular reasoning process is to specify the form of the memory structure.

The first empirical question addressed in the present research is, do juror's mental representations of the evidence assume a form that is consistent with the structure of a well-formed story? A second key empirical question concerns the extent to which a list of attributes serves as a valid model of the structure of information about verdicts in the juror's memory. One goal of the present research is to examine the functional significance of the proposed processing stages in relation to juror's predeliberation verdict decisions with a focus on the story construction stage. More specifically, a third empirical hypothesis is concerned with whether or not distinct story structures are associated with decisions for each verdict. If jurors favoring each verdict share a common mental representation of the evidence that is different from the stories affirmed by jurors favoring other verdicts, we will have established a correlation that is initial support for a conclusion that evidence structures cause verdict decisions.

An interview method was selected to elicit discussion of trial evidence and verdict definitions from jurors after watching a trial. The contents of their remarks were then described using a vocabulary adopted from current theories of causal reasoning and narrative discourse comprehension and from current theories of categorization. The induced structures were then compared for jurors choosing different verdicts. A necessary condition for demonstrating a functional relation between one or more of the proposed processing stages and the jurors' predeliberation verdicts would be to show a high degree of correspondence between variability in the induced structure and variability in jurors' verdict choices.

The sequence of steps in the present research can be compared to the sequence of studies reported by Bower and his students (Bower, Black, & Turner, 1979) in their investigation of script-based memory processing. In Bower's research, descriptions of common event sequences such as "going to the doctor" were elicited from subjects, and content analyzed. The overlap in these descriptions indicated that stereotyped event sequences, containing features of the proposed memory structures, were shared by subjects. Specific script structures were induced from these descriptions, and predictions for performance in memory and comprehension tasks were derived on the hypothesis that memory processing is script based. In our
research, we propose that episodic story schemata are the guides to the comprehension of trial evidence and its organization in memory. Furthermore, this organization of the evidence is functional in the decision-making stage of the juror's task.

Method

Subjects

Twenty-six adult subjects were randomly sampled from a pool of more than 200 volunteers in the Massachusetts Superior Court jury pool, during January and February of 1980. The subjects, 12 women and 14 men, ranged in age from 21 years to 73 years (average 47). These jurors had served on an average of three juries prior to the study and included people from all occupational and educational levels. This sample differed slightly from the jury pool from which they were drawn in containing a lower proportion of professionals and a lower proportion of women (see Pennington, 1981, for details). Jurors were not paid for their participation in the study because they received a wage for their service as jurors during the month.

Procedure

Four jurors were drawn at random each day from the group who had volunteered for participation in the study. The 4 subjects were taken to a jury room in the courthouse where they watched a videotaped trial. Prior to viewing the trial, subjects were instructed that an actual jury had decided the case and that their task was to "be one of the jurors" and to try to reach a decision after viewing the trial.

After watching the trial, each juror was interviewed individually; each was asked to talk out loud about the case and his or her decision while reaching an individual predeliberation verdict decision. There were two interviewers, so half of the subjects were interviewed immediately after viewing the trial, and the other half of the subjects were interviewed 1 1/2 hr later. The tape-recorded interviews were transcribed and submitted to extensive content analysis.

Stimulus Trial

The stimulus trial was a 3-hr videotaped reenactment of an actual murder trial in which the evidentiary information was maintained as accurately as possible. In the reenactment the judge and attorneys' roles were played by a practicing judge and two practicing trial attorneys. The witnesses were played by actors. This particular case and videotape has been used extensively in previous research on jury decision making. (Hastie, Penrod, & Pennington, 1983). The trial is involving and meaningful to actual jurors and has been rated by judges and lawyers to be realistic and representative of serious felony trials. (The reenacted trial, Commonwealth of Massachusetts v. Johnson, is summarized in the Appendix.)

In this trial, the juror chose from among four verdict alternatives in reaching a decision. The defendant Frank Johnson might be judged not guilty, guilty of manslaughter, guilty of second-degree murder, or guilty of first-degree murder. The jurors did not know at the outset of the trial that the four verdict alternatives were available. They knew only (as is usual) that the defendant was charged with first-degree murder.

The Interview

The major source of data in the present research was a protocol obtained in an interview with each experimental juror asking them to talk aloud about the Commonwealth of Massachusetts v. Johnson trial, and then to respond to questions about the evidence and about the judge's instructions to the jurors. The interview typically lasted about 1 hr and consisted of six parts. For present purposes, only two parts of the interview data are considered: the first part in which jurors spontaneously offered comments while thinking about a decision and the fifth part in which jurors were probed for their memory regarding the verdict definitions. In the first part, the juror was instructed, "I would like for you to be a juror in this case and to decide on a verdict. You do not need to decide right away. I would like for you to talk out loud as you think about what to decide." Experimenters probes in this part were restricted to statements urging the person to talk aloud while thinking. Subjects typically talked spontaneously for about 20 min. In the fifth part, probes were presented to elicit jurors' explicit reasoning about why they did not choose each of the alternate verdicts and to probe jurors' memory for the trial judge's instructions. Questions in this part were of the following form: "You have chosen (verdict chosen) as your verdict. How do you know it wasn't (alternate verdict)? What did the judge say (alternate verdict) was?" This part of the interview usually lasted about 10 min. (See Pennington, 1981, for details regarding other parts of the interview.)

We standardized interviewer behavior through training and assessment procedures. Two interviewers conducted all of the interviews in the research. Both interviewers were trained by conducting the interview with several pretest subjects and by observing one another. Additionally, a content analysis of interviewer behavior was conducted by coding each interviewer entry in the transcribed verbal protocols according to content and type of utterance. From this, a description of the "typical" interview content was produced, and analyses were conducted to examine possible differences between interviewers. There were no statistically reliable differences between interviewer behavior for immediate and delayed interviews, or for interviews of experimental jurors choosing different verdicts. However, one interviewer did pose approximately two more "contentless probes" (e.g., "Do you want to add anything to that?") per interview than the other (p < .002).

Results

Protocols were subjected to extensive content analyses. First, each experimental juror's predeliberation decision was identified. Second, protocol content was segmented into high-level content areas, for example separating story remarks and verdict remarks from other kinds of discussion content. Third, the story content was coded and transformed into a graph designating interrelations between events. The structure and content of these graphs were compared among different verdict decision groups. Next, verdict representation was coded and compared among different verdict groups. Finally, data that examined classification procedures were analyzed to determine the relation between verdict and decision criterion (standard of proof).

Decisions

Interview protocols were divided into groups according to each experimental juror's predeliberation decision. The verdict distribution across the 26 subjects was 8 guilty of first-degree murder, 9 guilty of second-degree murder, 6 guilty of manslaughter, and 3 not guilty (self defense). Sixteen of these subjects were chosen for analysis on the basis of the verdict chosen and the rated persuasiveness of the subject's verbal protocol, so that an even distribution of verdict choices and a diverse sample of reasoners within each verdict category were available for analysis. The persuasiveness ratings were an average of five judges' rank orderings of the protocols within each verdict category according to the overall persuasiveness of the arguments.
for the verdict chosen. The sample of 16 subjects contained 5 first-degree murder, 4 second-degree murder, 4 manslaughter, and 3 not guilty experimental jurors that included the most and least persuasive jurors in each decision category.

*Protocol Content Analyses*

The verbal protocol for each subject is a text that is the product of processing both the original trial information and reprocessing of the memory representation(s) of that information, influenced by the demands of our decision task. Even if the subject has spontaneously constructed a representation of trial information that resembles a narrative sequence, the protocols will not be direct replicas of such a structure (Kintsch & van Dijk, 1978). The content of the protocol text will contain elements relevant to the possible story constructed from evidence, but will also contain other evaluative, explanatory, and self-reflective statements embedded in a range of topics.

The first analysis applied to the protocol texts separated content into one of six different classes, two of which are relevant to our theoretical analysis: (a) *Story* statements include all references to events on the day of the crime-related activities of the defendant, victim, witnesses, and possible other participants. These are not references to events at the trial, but to events during the sequence of events testified to at the trial. For example, in the statement, “[1] The policeman said [2] the car was in front of the furniture store,” segment [1] is considered a *trial* reference because it refers to the event of the policeman testifying that occurred at the trial. Segment [2] is considered a *story* reference because the car’s location in front of the furniture store is an event that occurred on the day of the crime, not at the trial. In addition to explicit references to story events, this content category includes evaluations of story events, explanations and general world knowledge comments about story events, importance evaluations and other reflections on the subject’s own cognitions about the story (69% of the protocol). (b) *Verdict* statements include references to the trial judge’s instructions, verdict preference statements, elaborations of verdict definitions, requests for a rereading of portions of the judge’s instructions, and self-reflections related to verdict decisions (10% of the protocol).

Within the story content category, four subdivisions were made: explicit story reference, evaluative/hypothetical references, explanatory statements, and self-reflective statements. The first category is the actual story substance or content; the other three subcategories qualify as "metastatements" (Kintsch & van Dijk, 1978) or statements about the story. Two of these categories are of interest here: *Explicit story* references are actions, states, goals of characters, and enduring characteristics or habitual behavior of characters that are claimed to have occurred or not to have occurred within the context of the events referenced by the evidence. However, these need not be activities or states directly referred to in testimony; they are often inferred events. The explicit story references compose about 70% of the story content or about 50% of the total spontaneous protocol content and constitute the body of data that is analyzed in the greatest detail to induce story content and structure. *Story explanations* are statements of general truths about what generally happens in the world, statements of world knowledge invoked by the subject to explain or support conclusions about what actually happened in the story, and to assist in weighing plausible story events. These may take the form of the subject’s personal experience, general attitudes or beliefs, hypothetical self-analogy, or contrary-to-fact reasoning. Some examples are “If someone were going to fight with me, I wouldn’t go anywhere near them;” “Who wouldn’t be afraid?” “Sometimes when people drink they get nasty;” “Normally a person wouldn’t carry a big knife in their (sic) front pocket.” This category of remarks indicates the source of the juror’s story, that is, why certain interpretations of the evidence were believed to be true or plausible and other interpretations were rejected (Pennington & Hastie, 1980). About 11% of the story content was devoted to this kind of reasoning about the story.

Further consideration of story content in this section is restricted to coding of the explicit story references. In a later section, we will return to the story explanation remarks. The goal of the explicit story reference coding was to identify the events mentioned by the subject and to identify any connections among events asserted by the subject while talking about the case and the decision. The goal of this analysis was to pull out of the protocol a network of relations between events asserted by the subject.

First, within the explicit story reference category of protocol content, protocol text was coded into propositional units, defined as containing a single predicate relation (action or stative verb). Next, each proposition was assigned a specific content code from a list that included events referred to in testimony plus events frequently inferred by subjects. Next, the subject’s stance toward the truth of the proposition was coded as affirmed, neutral, or denied, along with a qualitative indication of certainty (certain or probable). Finally, all causal and temporal links between events asserted by the subject were coded.

A vocabulary of event types and causal links was developed based on Schank’s conceptual dependency theory of comprehension (Schank, 1972, 1975; Schank & Abelson, 1977; see also Graesser, 1981, for a similar system). Each of the proposition content codes was assigned an event type (goal, act, mental state, physical state, habitual state, or habitual act). Relations among events asserted by the subject were coded as links among events, using a vocabulary of causal and temporal relations that could link events (initiate, reason, result, enable, and then). Some constraints are imposed on what link types can connect which event types (Graesser & Goodman, 1985; Pennington, 1981). Two coders coded the protocols achieving levels of agreement ranging from 93% to 82% for different parts of the coding system.

The coding schemes resulted in a list for each subject of propositions with a content code, a stance, and links to other propositions. From this list, a directed graph structure was created for each subject in which the points represent event codes and the connecting arcs represent the links among events (Goodman & Hedetniemi, 1979). The network was further refined

---

1 Quantitative expressions of uncertainty were never used spontaneously by subjects in discussions of story events or verdict alternatives. The three uncertainty categories (certain, probable, neutral) were adequate to capture major differences in attitude toward the truth of a proposition and could be reliably applied.
by separating out that portion consisting of events and links affirmed as true or plausible by the subject. This structure, called a conceptual graph (Graesser & Goodman, 1985), captures part of each subject's representation of the evidence as indicated by the protocol events mentioned and assertions regarding relations among events. Two analyses of these networks are presented in the following section: (a) an analysis of the events and links common to subjects choosing each verdict category and (b) a graph theoretic analysis of central events for each verdict group.

Structure of Evidence for Jurors in Each Verdict Group

For the first analysis, a measure of shared features (Tversky, 1977; Tversky & Gati, 1978) was used to develop a central story for each verdict group. For each verdict group and for all subjects together, a group network was assembled containing only those event codes and links shared in common by 80% of the members of the group. This is a stricter criterion than has been used in some research examining common features of prototypical script representations (e.g., Bower et al., 1979). However, given the small sample, the 80% criterion resulted in an all-but-one rule for the verdict groups separately and an all-but-three rule for the total sample.

The event chains representing the common story structure for each verdict group and for all subjects combined are shown in Figures 2 through 6. An episode structure has been imposed by applying a series of rules to identify explicit goals linked to actions leading to the final consequence (Pennington, 1981). The shared central event chain identified in this way is summarized in story form for each verdict group below:

[First-degree murder story.] In the afternoon Johnson and Caldwell were in Gleason's Bar when a woman named Sandra Lee asked Johnson to give her a ride somewhere the next day. Caldwell became enraged because Sandra Lee was his girlfriend. So, Caldwell pulled a razor on Johnson and threatened him. Johnson did not have a weapon on him and went home. Johnson [angry or humiliated] decided to confront Caldwell and show him once and for all who was boss, by killing him if it came to that. Later Johnson's friend Clemens came over and agreed to back him up. Johnson got his knife and Clemens and Johnson went back to the bar to find Caldwell. They went into the bar and after awhile, Johnson and Caldwell decided to go outside and have it out. Caldwell hit Johnson, renewing Johnson's vow to get Caldwell once and for all. He pulled his knife, raising it up, and thrust it into Caldwell. A policeman saw the stabbing and rushed over but Caldwell died almost immediately from a massive chest wound.

[Second-degree murder story.] Johnson had an argument with Caldwell in Gleason's Bar in the afternoon. That evening Johnson and his friend Clemens decided to go back to Gleason's Bar, where they were regulars. Johnson was quite certain that there would be more trouble with Caldwell so he put his fishing knife in his pocket just in case. Johnson and Clemens went to Gleason's and it was not long before Caldwell and Johnson had resumed their argument, leading to a fight. A policeman across the street saw Caldwell and Johnson slug it out when Johnson pulled a knife from his pocket, raised it up over his head and viciously stabbed Caldwell in the chest. This resulted in a massive stab wound to Caldwell, killing him almost immediately. Later at the autopsy, Caldwell's weapon was found in his back pocket.

[Manslaughter story.] In Gleason's Bar in the afternoon, Caldwell threatened to kill Johnson so he left the bar. Then later that night, Johnson went back to the bar. After awhile he and Caldwell went outside to talk over their differences, so Johnson thought, having had a history of arguments with each other and a longstanding jealousy between them. Outside, Caldwell punched Johnson, alarming him as he remembered Caldwell's earlier threat. Johnson pulled a fishing knife that he was carrying, intending to prevent Caldwell from harming him with a razor as he had previously threatened. In the confusion of the fight he stabbed Caldwell, but later, Caldwell's razor was found in his back pocket.

[Not guilty story.] Caldwell and Johnson were in Gleason's Bar one afternoon when a woman named Sandra Lee asked Johnson for a ride somewhere the next day. Caldwell suddenly became violent, pulled out a razor and threatened to kill Johnson. Thinking it best to get out of Caldwell's way, Johnson left the bar. Later that evening, Johnson's friend Clemens came over to Johnson's house and suggested that they go to Gleason's for a beer. Johnson didn't want to go but agreed to, if they checked first and Caldwell wasn't there. At the bar, they saw that Caldwell was not there and so they went in. However, Caldwell came in later and asked Johnson to step outside. As soon as they went outside, Caldwell punched Johnson unexpectedly, knocking him back hard against the wall of the building. Johnson was woozy from the punch and he knew that Caldwell had a razor because Caldwell had threatened him with it earlier. Johnson was carrying his fishing knife, which he usually carried, so he pulled it out to show the knife and keep Caldwell from coming after him. As he came back from the wall he saw that Caldwell had his razor out and was coming at him. Remembering Caldwell's earlier threats, he thought he had better protect himself so he held his knife out in front of him. Either Caldwell lunged into it, or as Johnson put the knife out, he thrust it a little and the knife went into Caldwell, wounding him.

[All subjects' story.] Johnson and his friend Clemens went to Gleason's bar in the evening. Another man, Caldwell, punched Johnson so Johnson pulled his knife and stabbed him.

Figure 2. Central story for mock jurors choosing first-degree murder. (J = Johnson; Ca = Caldwell; S.L. = Sandra Lee; Cl = Clemens.)
The stories also have a beginning, middle, and end in typical narrative style, and the story structures can be represented as hierarchies of embedded episodes. For example, in the first-degree murder story the events included in the box labeled as initiating events (see Figure 2) form an embedded episode that can be referred to as the "argument in the afternoon." In these analyses, the highest level episode for each verdict group's central story was determined by examining the shared story network for the verdict group and identifying the goal with the direct causal chain to the final consequence. The highest level episode should reflect those events and goals that are the core of the story for the juror, that is, those that are the most important. To check this, a second analysis determined the central causal chain for each juror separately. A measure of centrality was developed by first deriving a matrix of distances between each pair of story events for a subject. The distances were the number of links in the shortest path between two events in the subject's conceptual graph. Central events were defined as those with the smallest average distances to and from them across all other events. The central causal chain for each subject was determined by including only these core events and the causal links among them. Aggregating across the central causal chains for jurors within each verdict group, an 80% criterion, now yields a much abbreviated shared verdict story. However, this analysis reflects the episode hierarchy of the first analysis in that the ver-
The experimental juror began her protocol in this manner, attempting to identify the “to be proven” and then to systematically list features of inductive support. This juror soon abandoned this reasoning strategy as too confusing. She subsequently reverted to causal reasoning and narrative form to summarize her remarks. A third possibility is that the important evidence revolves around the characterizations built of the defendant and victim so that structures emerging from protocols would show partitions of character sketches connected to verdicts. Experimental jurors did make character inferences (5.4% of story content), but these were integrated into the story structures as reasons for certain behaviors. For example, an inference that Caldwell was a violent man might be given as a reason that Caldwell pulled a razor when provocation was slight or as a reason that Johnson was afraid.

A second major conclusion from these analyses is that story structure and content are different for different verdict groups. Although this can be seen from Figures 2 through 6, a statistical summary of this finding is presented in Table 3. For each juror a similarity index between the juror’s story structure and each of the four verdict story structures was computed by subtracting the proportion of nonverdict story events included in the juror's story (mismatches) from the proportion of verdict story events included in the juror's story (matches). This number can range

![Diagram](image-url)
from -100% to +100%. A negative similarity score indicates that a greater proportion of nonverdict story events than verdict story events were included in the story. Jurors' stories are clearly more similar to their chosen verdict story than to the stories for other verdict categories as shown by the interaction between verdict story and verdict choice in Table 3, F(6, 26) = 15.34, p < .001. This demonstrates that the induced verdict category story structures are reliably different for the jurors choosing different verdicts.

In addition to finding differences between verdict stories, the features of the main episodes map directly onto the required features for the respective verdict categories. For example, in the not-guilty story (see Figure 5), there are three episodes, two of which are embedded. The main episode is the fight, and two events in the main episode contain other episodes. Going into the bar is the outcome (consequence) of a highly elaborated episode initiated by Clemens and containing its own goal-directed behavior. The consequence is that Johnson was at the bar, his physical state resulting from habit and not directed by goals relevant to the main episode. The initiating events are all Caldwell's actions during the fight. The entire afternoon episode acts to fortify not guilty subjects' conclusions about Johnson's psychological state, leading first to a goal to show the knife and then to actively protect himself. The not guilty story shows the knife going into Caldwell as a consequence rather than as a goal-directed action. There features correspond to the verdict features of not guilty by reason of self defense: under immediate attack, unable to escape, intent to defend, and reasonable retaliation. In contrast, first-degree murder requires premeditation, that is, a resolution formed to kill, an interval of time, and a killing in accordance with the resolution. The subjects' stories (see Figure 2) emphasize initiation of an intent to kill goal and the causal relatedness of subsequent actions leading to the stabbing and Caldwell's death. The common story across all subjects combined (see Figure 6) is incomplete with respect to the episode schema and with respect to any verdict category. The major components missing in the common story are the psychological states, goals, and consequences. Because essential, interpretive components are missing from the episode, the story fragment could potentially fit any of the four verdict categories.

These events form the focal point of the evidence to be explained. However, the explanation, the interpretation constructed around these events, is not shared across subjects.

To summarize, two major conclusions follow from these analyses. First, experimental jurors' representations of evidence showed story structures and not other plausible structures. Second, jurors who choose different verdicts have different interpretations of the evidence, that is, different stories. Furthermore, differences in stories correspond to differences in verdict definitions suggesting a functional role for representations of evidence in this decision task.

**Verdict Category Representations**

There is considerable variation among jurors in the accuracy and completeness of their representations of the verdict category information. If these variations are systematically related to the jurors' verdict choices, then the hypothesis has to be retained that differences occurring in the verdict category establishment stage also influence the verdict choice. For example, if jurors favoring first-degree murder verdicts were also jurors who did not remember the judge's instruction concerning premeditation, it would be plausible that the memory failure affected the verdict choice.

The interview schedule included several items that inquired into the juror's knowledge of information presented by the trial judge concerning the verdict categories. These questions were of the form, "Do you remember how the judge described second-degree murder?" A coding scheme based on the features for each verdict category listed in Table 1 was applied to responses to these items. A prototype response to one of these items would include three terms: reference to a verdict category (e.g., first-degree murder), reference to a crime element (e.g., premeditation), and a relation between the two (e.g., first-degree murder includes premeditation). Some responses by subjects did approximate this ideal, for example, "First-degree murder had

---

2 Because verdict story structures were induced from subject protocols, the match part of the similarity score will necessarily be high for the juror's chosen-verdict story. However, the match proportions for the other-verdict stories and the mismatch proportions will not be low unless the verdict story structures differ from each other.
malice and it was premeditated.” Other responses were not as well formed, for example, “First-degree murder, that’s premeditated. It doesn’t matter whether I plan it for five minutes or ten hours, it’s planned. You know you’re going to do it before you do it.” The coding scheme was applied in a lenient manner; if all three terms of a relation were explicit or required little inference on the part of the coder, the juror was credited with having cited the element-verdict relation (87% agreement was obtained between two coders).

The coding categories departed from the set presented in Table 1 in two ways. First, the “identity” element listed in Table 1 was not considered in the coding because all respondents took it for granted that Johnson killed Caldwell given the uncontested fact pattern on this issue. Second, some of the elements were combined (e.g., formed a plan and killed in accordance with the plan) into a single coding feature. In the final scheme, there were 10 features for the first-degree murder verdict, 5 for second-degree murder, 5 for manslaughter, and 6 for not guilty.

The results of the analysis of jurors’ verdict category representations are summarized in Table 4. Subjects made very few errors of commission, improperly relating an element to a verdict category (less than 10% of the total responses), and no juror made more than two errors of commission. The low frequency of errors (10) precludes systematic analysis. One observation is that the second-degree murder verdict was associated with the largest number of errors. The remaining analyses will exclude the 10 erroneous responses.

The data in Table 4 are cross-classified by verdict category (e.g., all references to the first-degree murder verdict and its elements) and juror verdict choice (e.g., all jurors who selected the first-degree murder verdict). A clear pattern emerges when an analysis of variance (ANOVA) is computed on the total numbers of verdict-element citations for verdict category or verdict-choice classifications. More citations were made to the first-degree murder verdict category, \( F(3, 36) = 4.47, p < .01 \), than to the other three verdict categories. The second factor, juror verdict choice, did not have a significant effect on citation of verdict-element relations, \( F(3, 12) = .40 \). Other analyses of specific verdict elements by verdict choice showed no content differences. Thus, there is no evidence for a systematic relation between verdict category representation and verdict choice. Variance in performance in the verdict category establishment stage of the juror decision does not appear to affect the jurors’ decision.

### Story Classification

The interview used in the present research did not obtain information concerning the juror’s notions of presumption of innocence or the beyond reasonable doubt standard. However, other research sampling jurors from the same population of jurors and using the same stimulus trial (Hastie et al., 1983) did obtain direct ratings of subjects’ subjective values for the beyond reasonable doubt threshold and the presumption of innocence principle. In that research, two questions were put to 81 jurors from Superior Court jury pools in Massachusetts after viewing the *Commonwealth of Massachusetts v. Johnson* trial and rendering individual verdicts on that trial.

One question asked about presumption of innocence, “In terms of probability, what degree of certainty concerning the defendant’s innocence should you feel before any evidence is presented at trial relevant to his guilt or innocence?” A second question asked about the reasonable doubt standard, “In terms of probability, to what degree of certainty do you feel you must be convinced that the defendant committed the alleged offense before you find him guilty?” Subjects indicated their responses on a 100-point scale labeled 0%, 10%, 20%, ..., 100%.

An ANOVA conducted on these ratings did not find significant differences among the subjects when they were classified by verdict choice (not guilty, 87%; manslaughter, 89%; second-degree murder, 91%; first-degree murder, 89%). The implication for the present work is that variation in performance in the story classification stage is not associated with variations in verdict choice.

### Summary of Results

A three-stage Story Model of juror decision making has been proposed that includes story construction, verdict category establishment, and story classification stages. The empirical research has focused on the causal structure of jurors’ stories and its relation to verdict choice. A minimum level of support for the model requires (a) that the contents of jurors’ spontaneous protocols exhibit organizational characteristics such as causal chains and episodes that are consistent with narrative story structures, and (b) that jurors who choose different verdicts tell different stories.

The first major empirical finding is that spontaneous interview protocols do exhibit story structures. When subjects were asked to talk aloud about their decisions, content coding of the protocol text yielded individual and verdict group structures that have causal event chain and episode structures. Juror’s stories were not simple lists of evidence. They always contained most components of the episode schema (initiating events, psychological and physical states, goals, actions, and consequences) in appropriate causal relations. Jurors typically inferred missing components for these structures when they were not contained in direct testimony at trial. Evidence not related to a story of what happened was systematically deleted from discussion. In addition, jurors’ stories contained events and goals that
were sensibly related to verdict category features. Specifically, psychological states and goals of actors in the stories were related to features of the verdict categories designated as legal elements of the charge against the defendant.

The second major empirical question, concerning the categorical representation of verdict alternatives as feature lists, was also answered in the affirmative. Subjects tended to provide lists of crime elements in a fashion similar to the lists of attributes that have been elicited as evidence of feature list representations for natural language noun categories (Rosc'h & Mervis, 1975). However, subjects' category representations were far from complete as indicated by the average number of citations compared to the legal definitions.

The third major empirical finding is that stories vary systematically by verdict choice both in content and structure. Content differences appeared in goals and mental states said to cause behavior of the defendant and other actors. These content differences corresponded to specific, legally required features of the jurors' chosen verdict categories. More important, central stories for the different verdict groups differed radically in causal event chain and episode structure. Thus, the organization of evidence differed for subjects choosing different verdicts. Shared causal event chains and episode structures for different verdict groups showed differential elaboration and different causal link patterns that mapped directly onto verdict features. Two different analyses identifying a central story for each subject and the shared central story for each verdict group supported the episode structure inferred from the shared causal event chains. Higher level verdict group summaries were, again, related to verdict choice.

The fourth major empirical finding is that verdict category structure and story classification procedures do not vary systematically by verdict group. Subjects' representations of verdict categories, compared across verdict choice groups, did not differ systematically. Therefore variability in verdict choice is not accounted for by different representations of verdict category meaning. Although the story classification stage was not examined in detail in this research, data from another study using the same stimulus trial (Hastie et al., 1983) suggest that the presumption of innocence and standard of proof criteria also do not vary systematically according to the verdict choice of the subject. Thus, the story classification procedures are also ruled out as the locus of verdict-choice differences.

Discussion

The proposed Story Model of juror decision-making provides a conceptual model for what jurors are actually doing when asked to decide on a verdict for criminal cases. The Story Model divides the decision process into three stages: (a) The story construction stage specifies processing that includes the evaluation of the meaning and relevance of stimuli. Meaning is assigned to trial evidence through the incorporation of that evidence into one or more plausible accounts or scenarios describing what happened during crime-related events testified to during the trial. In addition, general knowledge about the structure of human action sequences characterized as an episode schema serves to organize events according to the causal and intentional relations among them as constructed by the juror. (b) The verdict category establishment stage specifies the legal information as to what constitutes medical malpractice, yet the presence or absence of medical malpractice was the main legal issue to be decided. Fourth, the multidimensional scaling analysis, based on a Euclidean spatial model of relations between items, used in the Ryan and Hurtig research, is not an appropriate model to represent causal relations between items. In our research, subjects, setting, stimulus case, instructions, and task are much closer to actual jury trial condi-
tions than those in the Ryan and Hurtig research. In addition, our analyses were designed to capture the structure of mental representation in a manner commensurate with plausible, theoretical expectations about the natural forms such structures might assume.

One question that arises is how the story constructions come to differ among jurors. According to our theory, stories are constructed by reasoning from world knowledge and from evidence. Some evidentiary items are established as true directly on the basis of their appearance as evidence from one or more credible sources. Which of these events will appear as relevant depends on the interpretation assigned to the facts from their causal relatedness to other events inferred and testified to. The inclusion of other evidentiary items, inferred events, and relations among them are the result of a wide variety of deductive and inductive reasoning procedures applied to the evidence and world knowledge.

Analyses of inference chains establishing story events reveal that intermediate conclusions are established by converging lines of reasoning that rely primarily on deduction from world knowledge, analogies to experienced and hypothetical episodes, and reasoning by contradiction (Pennington & Hastie, 1980). Examples are instructive in illustrating sources of uncertainty in story construction.

A typical deduction from world knowledge consists of the following premise (P1-P3) and conclusion (C) structure: P1 = A person who is big and known to be a troublemaker causes people to be afraid. P2 = Caldwell was big. P3 = Caldwell was known to be a troublemaker. C = Johnson was afraid. In this example the juror matches features of Caldwell from evidence (P2) and from a previous inferential conclusion (P3) to world knowledge about the consequences of being confronted with such person (P1) to infer that Johnson was afraid (C). On the surface this kind of inference looks like a representativeness judgment (Tversky & Kahneman, 1983) in that the features of the concrete situation (P2 and P3) are matched to the features of a prototypic situation (P1), and the certainty associated with the conclusion should be related to the goodness of the match. Tversky and Kahneman (1974) have provided evidence that the goodness of match will influence the certainty of the conclusion, but we believe that the judgment heuristics include only a subset of the conditions that affect certainty assessment. A more complete treatment is available in Collins's theory of plausible inference (Collins, 1979).

Collins proposes a set of inference procedures in which the traditional rules of logic are generalized from two values of certainty to a continuum of degrees of certainty. Each inference procedure is associated with a set of "certainty conditions." For example, for a deductive inference procedure, certainty conditions would include the following: (a) typicality of the instances of their categories (e.g., How typical a bully is Caldwell? How typical a person is Johnson? Note, this is Kahneman & Tversky's "representativeness"); (b) baserate of the properties under consideration (e.g., How often are people afraid of other people?); (c) variability in the knowledge from which the general premise arises (e.g., Is there a wide range of reactions to bullies?); (d) dissimilarity between the instance and the most similar group that behaves in a discrepant manner (e.g., How dissimilar is Johnson from the kind of person who may have a different reaction to a bully?). These conditions are developed more fully in Collins (1979), and empirical support for parts of the plausible inference system have been provided by Rips (1975) and Collins (1978; Collins, Warnock, Aiello, & Miller, 1975).

Confidence in the conclusion of an inference is also assessed by reasoning by analogy to other experiences and by evaluating alternate conclusions that would contradict the initial conclusion. For example, the same juror who provided the premise--conclusion example just mentioned, continued with, "If someone like Caldwell came up to me in a bar and threatened me, I would be afraid." Alternate reactions were also considered, "I don't think Johnson was angry. If he had been angry, he would have gone right back to the bar. He didn't go right back." And some were rejected, "No, Johnson was afraid of Caldwell and he took his knife with him because he was afraid."

A juror with different attitudes, experiences, and beliefs about the social world, would reach different conclusions. For example, another juror reaches the opposite conclusion using virtually identical reasoning procedures. "Caldwell embarrassed Johnson. I'm quick-tempered. If someone embarrassed me, I would go back with malice and do a job on him. Johnson is quick-tempered, too. He went back with malice to do a job on Caldwell." Here is another example of hypothetical reasoning by analogy to the self to infer internal reactions of a character. This juror also checks for consistency with the evolving story, "If Johnson were afraid, he would have stayed home, and we know he didn't do that."

To summarize, the story is established by inferring events not included in the testimony from frames of world knowledge matching events already established. Inferences are evaluated by simulating one's own behavior in similar situations, by checking for contradictions with other plausible conclusions, and by checking for inconsistencies with the current form of the story. These inference chains are often complex and exhibit a variety of reasoning strategies (Pennington & Hastie, 1980). In juror decision making, it is almost never the case that an inference chain is connected directly to assessments of guilt. Rather, the story is constructed piece by piece from the evidence, and the story as a whole is used to choose a verdict.

A second question concerns the temporal relations among Story Model processing stages. Ryan and Hurtig (1980) raised the question of whether or not the structure of the evidence base "causes" verdicts. Their study failed to find a relation between variance in the structure of the evidence base and variance in decisions. Because such a relation is a necessary condition to conclude that the structure assigned to the evidence "causes" the decision, they rejected the hypothesis. In contrast, the current study found a relation between evidence base structure and verdict decision. However, the question of a direct causal relation between stories and verdicts cannot yet be answered definitively. Both studies are correlational; the fact that a relation between the story constructions and verdict decisions of jurors has been found provides some support for the proposed Story Model. However, there are other possibilities for the "causal" relation: (a) Jurs' stories are instrumental in reaching a decision, or (b) jurors' decisions compel them to construct stories and explanations to justify the decision, or (c) there is an interaction between the construction of stories and the decision pro-
The assignment of meaning and the elaboration of event sequence representations enable jurors to reach appropriate linguistic/comprehension models of causal reasoning. In conflict decisions. The proposed model is representative of current computations of judgment on a single dimension. The present research suggests that moving the focus of decision making investigations to the interpretation component rather than the computational component of the decision task may change one's conception of the processes that occur during the computation. 

This research represents a broader perspective on decision making that we have come to think of as explanation-based decision making. Many naturally occurring decisions begin with a review of relevant evidence to construct a causal model comprising key events and relations among them. Researchers studying medical diagnosis (Pope, 1982), engineering (Brown, Burton, & de Kleer, 1982), diplomatic negotiation (Axelrod, 1976), business forecasting (Einhorn & Hogarth, 1982), and other tasks have also placed causal reasoning in a primary role. However, domain-specific causal rules and structures are found in the different spheres of activity; an internist's causal model of a patient's condition and its predecessors will not be represented in the same mental concepts as an engineer's mental model of a troublesome electrical circuit or a juror's story of the crime and the events surrounding it. Nonetheless, experts in each domain will begin their decision process by constructing a causal model to explain the available facts. It is also likely that the decision will involve a category classification process analogous to verdict selection stage in the juror's Story Model and that confidence in the final decision will be determined by the coherence of the causal model of the facts and the goodness-of-fit of the model to the selected decision category. 

The role of causal reasoning in explanation-based decisions is distinct from traditional social psychologists' analysis of causal attribution processes in two respects. First, following Brown and Fish (1983) and Schank and Abelson (1977) we have described the causal reasoning occurring as the linguistic operations involved in constructing a knowledge structure representing some part of the real world. The emphasis on comprehension operations with a deductive character is a contrast to the inductive calculations associated with attribution models posited by Kelley (1973), Jones and Davis (1965), and their colleagues. Second, most research on causal attribution has studied decisions in which a judgment of a cause's identity or strength defines the subject's task. In the juror's task, and the other examples we have cited, causal attributions are instrumental in the decision task, serving the purpose of identifying relevant information through comprehension processes.

References


Kahan, Law & Cognition


Caldwell and Johnson didn’t come to the bar together. First, Johnson ran to the window because there were some shouts outside. Actually, bar and then they walked outside together. On the way out Caldwell put Caldwell arrived. Then later, Caldwell and Johnson were talking at the at an angle. Frank Johnson has a reputation for peacefulness and has both been in the bar. At that time they were arguing. Caldwell pulled both been in the bar. At that time they were arguing. Caldwell pulled his razor on me so I stuck him.” (This last statement was declared inadmissible by the trial judge.) The knife Harris retrieved from the ground near Caldwell is introduced as evidence. It measures eleven inches from end to end.

State Pathologist, Dr. Robert Katz

I found the following items on the body of Alan Caldwell: a ring, a watch, and small change in the right front pocket, and a straight razor in the left rear pocket. Caldwell was killed by a stab wound to the heart between the third and fourth ribs. I was unable to determine the angle at which the knife entered Caldwell’s chest. His blood alcohol level was .032. This is enough alcohol that Caldwell may have been drunk. Caldwell had numerous surgical scars on his body. There were other scars of undetermined origin. The straight razor is introduced as evidence.

Patrick Gleason

I am the owner of Gleason’s Bar and Grill. That night I had occasion to run to the window because there were some shouts outside. Actually, I expected it because I had watched Caldwell and Johnson leave the bar together a few minutes before. Through the window I saw Johnson raise his hand up and stab Caldwell. I didn’t see anything in Caldwell’s hand although I could not see Caldwell’s right hand which was down by the side away from me. Johnson did not resist arrest, but he did say, “Caldwell pulled a razor on me so I stuck him.” (This last statement was declared inadmissible by the trial judge.) The knife Harris retrieved from the ground near Caldwell is introduced as evidence. It measures eleven inches from end to end.

The Defendant’s Plea

The defendant, Frank Johnson, pleads NOT GUILTY.

Officer Richard Harris

On May 21st at about 11:00 P.M. I was on my usual foot patrol when I heard shouts from the direction of Gleason’s Bar and Grill and hurried in that direction. From across the street I saw Caldwell (the victim) hit the defendant Johnson in the face. Johnson staggered back against the wall, then came forward and raised a knife above his head with his right hand. I yelled, “Frank, don’t do it.” But he plunged the knife downward into Caldwell’s chest. Caldwell had fallen to the ground by the time I reached the scene. I apprehended Johnson, phoned for a police cruiser and an ambulance.

Cross-examination. I had a clear view of the fight from across the street approximately 75 feet away. I did not see anything in Caldwell’s hand although I could not see Caldwell’s right hand which was down by the side away from me. Johnson did not resist arrest, but he did say, “Caldwell pulled a razor on me so I stuck him.” (This last statement was declared inadmissible by the trial judge.) The knife Harris retrieved from the ground near Caldwell is introduced as evidence. It measures eleven inches from end to end.

State Pathologist, Dr. Robert Katz

I found the following items on the body of Alan Caldwell: a ring, a watch, and small change in the right front pocket, and a straight razor in the left rear pocket. Caldwell was killed by a stab wound to the heart between the third and fourth ribs. I was unable to determine the angle at which the knife entered Caldwell’s chest. His blood alcohol level was .032. This is enough alcohol that Caldwell may have been drunk. Caldwell had numerous surgical scars on his body. There were other scars of undetermined origin. The straight razor is introduced as evidence.

Patrick Gleason

I am the owner of Gleason’s Bar and Grill. That night I had occasion to run to the window because there were some shouts outside. Actually, I expected it because I had watched Caldwell and Johnson leave the bar together a few minutes before. Through the window I saw Johnson raise his hand up and stab Caldwell. I didn’t see anything in Caldwell’s hand although I could not see Caldwell’s right hand which was down by the side away from me. Johnson did not resist arrest, but he did say, “Caldwell pulled a razor on me so I stuck him.” (This last statement was declared inadmissible by the trial judge.) The knife Harris retrieved from the ground near Caldwell is introduced as evidence. It measures eleven inches from end to end.

Caldwell and Johnson didn’t come to the bar together. First, Johnson ran to the window because there were some shouts outside. Actually, bar and then they walked outside together. On the way out Caldwell put Caldwell arrived. Then later, Caldwell and Johnson were talking at the at an angle. Frank Johnson has a reputation for peacefulness and has never caused trouble in the bar. (The judge does not allow Gleason to testify about Alan Caldwell’s reputation.)

Dennis Clemens

I stopped at Frank Johnson’s home on the evening of May 21st and asked Johnson to join me for a drink at Gleason’s, which is where we usually go. Before we went in the bar, we looked around. We didn’t see anything. At about 9:30 P.M. Caldwell entered, and after a while motioned Johnson to come and talk. In a few minutes, Johnson and Caldwell left the bar. I could not hear what they said, but went near the front door which was open. I heard a few shouts, saw Caldwell punch Johnson to the ground, and begin to attack him with a razor. Johnson tried to hold Caldwell off but Caldwell attacked, there was a scuffle, Caldwell staggered back, and after about twenty seconds fell to the ground. I didn’t go outside to stop the fight because it happened so quickly.

Cross-examination. Johnson and I did not go to Gleason’s looking for Caldwell, and Johnson was reluctant to go into Gleason’s until we had assured ourselves that Caldwell was not there. I saw the razor clearly in Caldwell’s right hand. I didn’t see the knife in Johnson’s hand because of the angle of the two men.

Janet Stewart

I am a waitress at Gleason’s Grill, and on the night of the fight I noticed both Caldwell and Johnson in the grill before the fight. There was shouting outside. When I ran outside I saw Caldwell on the ground. I also noticed Caldwell’s car, which I recognized, was parked illegally in front of the grill and would have obstructed a view of the fight from across the street.

Frank Johnson

I was in Gleason’s Grill on the afternoon of May 21st. A woman asked me to give her a ride somewhere the next day. Alan Caldwell immediately came over screaming at me and threatening me; he pulled a razor and threatened to kill me. I was quite upset and frightened and I went home and spent the day with my wife and six children until 9:00 P.M. when Dennis Clemens came by and suggested we go out for a drink. When we got to Gleason’s Grill, I was afraid to go in but was finally convinced when we could find no evidence that Caldwell was in the grill. Later Caldwell entered and sat at the bar. Twenty minutes later Caldwell motioned me over in a friendly way and suggested we go outside. Caldwell was calm and friendly and I thought he wanted to talk. Once outside though, Caldwell became angry, threatened to kill me, and then hit me in the face with his right fist. The blow knocked me back against the wall and stunned me but I noticed that Caldwell had pulled his razor again. I unthinkingly reached for my fishing knife, pulled it out and held it in front of me to keep Caldwell away. But Caldwell rushed in with his razor and lunged on the fishing knife. The next thing I remember is Officer Harris arriving at the scene. I almost always carry my fishing knife because I am an avid fisherman and my wife does not like the knife to be lying around where the smaller children may pick it up. I couldn’t get away from the fight because I was knocked down and cornered against the wall. I reached for the knife instinctively and tried to protect myself. I didn’t mean to kill Caldwell.

Cross-examination. I don’t think I had my knife with me in the afternoon but I don’t really know because I carry it with me a lot.
Judge’s Instructions

Now I want to define the elements of the charges against the defendant. These definitions are essential to determining which verdict you should return. The Commonwealth has charged the defendant with murder in the first degree. You actually have four verdict alternatives: murder in the first degree, murder in the second degree, manslaughter, and not guilty by reason of self defense. Murder in the first degree is a killing committed with deliberately premeditated malice aforethought. Both first and second degree murder require malice aforethought. This includes feelings of hatred or ill will, but is more than that. It means any intentional infliction of an injury where there is a likelihood of causing death. Malice may be inferred from the intentional use of a deadly weapon without just provocation or legal excuse. If you find a killing with malice, then you must return a verdict of at least murder in the second degree. To return a verdict of first degree murder, you must find that the killing was performed with deliberately premeditated malice aforethought. Deliberate premeditation is a sequence of thought processes. You must be convinced that a plan to murder was formed. It is the sequence of thought, not the time taken that determines premeditation; it may take place over weeks, days, hours, or only seconds. First the deliberation and premeditation, then the resolution to kill, then the killing in pursuance of the resolution.

The third verdict alternative is manslaughter. Manslaughter is a killing without malice, a killing resulting from a sudden passion or heat of blood produced by a reasonable provocation or sudden combat. Reasonable provocation requires more than just words or insults or a general dislike or general feeling of anger. Your final verdict alternative is self defense. If you find that the killing was in self defense then the defendant is not guilty of a crime and should be found “not guilty”. The right to self defense arises from a threat to one’s life but does not come into existence at all until the defendant has exhausted all reasonable means to escape from the confrontation and once the threat is over, the right evaporates. The method one uses to defend oneself can only be reasonable, but this judgment is made with some consideration for the frailties of human impulses in a stress situation. If the defendant does not have a reasonable fear of great bodily harm or has not exhausted all reasonable means of escape or has used more than reasonable force to protect himself, then self defense does not apply and the defendant is guilty of at least manslaughter. Since the defendant has raised the issue of self defense, the burden is on the Commonwealth to prove it was not a situation of self defense. Finally, remember it is your duty to find the defendant guilty of the more serious charge of which he is in fact guilty beyond a reasonable doubt. If the Commonwealth has failed to prove the elements of the charged offenses, it is your duty to find the defendant not guilty.

Received August 26, 1985
Revision received February 5, 1986
A COGNITIVE THEORY OF JUROR DECISION MAKING: THE STORY MODEL

Nancy Pennington and Reid Hastie*

INTRODUCTION

The goal of our research over the past ten years has been to develop a scientific description of the mind of the juror as it is revealed in the legal decision-making process. Our conclusion is that the juror is a sense-making information processor who strives to create a meaningful summary of the evidence available that explains what happened in the events depicted through witnesses, exhibits, and arguments at trial.

We begin the presentation of our views with a quick survey of the images of the juror that have been significant in modern legal decisions and scholarship. However, we could find no succinct summary of the assumed decision processes of a typical juror in the extensive literature emanating from law schools, courts, and other legal authorities. From clues and fragments appearing in rules of evidence, appellate decisions, and law texts, we can infer a common image of a "reasonable man" who is capable of rough-and-ready logical deductions, but who is also likely to be prejudiced, swayed, or diverted by sentiment-evoking evidence. The psychological literature also lacks any general unified discussion, although an unflattering image of the juror can be discerned in the multitude of references to a bias-prone creature who constructs a decision from a toolbox of prejudices and heuristics.

Probably the most unified descriptions of the juror's thought processes are mathematical models based on Bayesian probability theory, variants of traditional probability theory, and other algebraic

* Psychology Department, University of Colorado. Address all Correspondence to: Nancy Pennington, Psychology Department, Campus Box 345, University of Colorado, Boulder, Colorado 80309-0345.

Research reported in this paper was supported by the National Science Foundation, Law and Social Sciences Program. The authors would like to thank Ronald Allen and Richard Lempert for comments on a previous draft of the paper.


models. However, none of these approaches has dominated legal scholarship or practice and none has been supported as a descriptive model by empirical research on realistically complex juror decision-making tasks. So, the field seems open for a new descriptive model of juror decision making.

We have approached the goal of describing juror decision making with the perspective of psychologists who are interested in how people think and behave. First, we have developed a theory that we believe describes the cognitive strategies that jurors use. We call this theory the Story Model, and it is described in the first section of the paper. Second, we have conducted extensive empirical research to test the theory. This work is summarized in the second section of this paper. The Story Model, as is true for most theories, is not complete in its current form. In the last section of this paper, we discuss directions for future theory development and research.

I. THE STORY MODEL

We call our theory the Story Model because we propose that a central cognitive process in juror decision making is story construction. Although story construction is central in our theory, and has been the focus of most of our empirical research, it is but one of three processes that we propose. In overview, the Story Model includes

---

three component processes: (1) evidence evaluation through story construction, (2) representation of the decision alternatives by learning verdict category attributes, and (3) reaching a decision through the classification of the story into the best fitting verdict category (see Figure 1). In addition to descriptions of processing stages, one central claim of the model is that the story the juror constructs determines the juror's decision. As part of the theory, we also propose four certainty principles—coverage, coherence, uniqueness, and goodness-of-fit—that govern which story will be accepted, which decision will be selected, and the confidence or level of certainty with which a particular decision will be made.

In the next sections of the paper we describe the processing stages proposed in the Story Model and the certainty principles that govern them. In order to illustrate our ideas with examples, we will draw on one of the simulated trials that we have used in our research, Commonwealth of Massachusetts v. Johnson. In this trial, the defendant Frank Johnson is charged with first-degree murder. The undisputed background events include the following: the defendant, Johnson, and the victim, Alan Caldwell, had a quarrel early on the day of Caldwell's death. At that time, Caldwell threatened Johnson with a razor. Later in the evening, they were again at the same bar. They went outside together, got into a fight, and Johnson knifed Caldwell, resulting in Caldwell's death. The events under dispute include whether or not Caldwell pulled a razor in the evening fight, whether Johnson actively stabbed Caldwell or merely held his knife out to protect himself, how they got outside together, whether or not Johnson intentionally went home and got his knife, whether Johnson went back to the bar to find Caldwell or went to the bar because it was his habit, et cetera.

A. Constructing a Story

The Story Model is based on the hypothesis that jurors impose a narrative story organization on trial information. According to the theory, the story will be constructed from three types of knowledge

---

8 This is a simulated case. See infra note 10.
9 The Story Model has been developed in the context of criminal trials, so it will be presented and discussed in those terms. In the final section of the paper, we discuss its extension to civil trials. See infra p. 551.
10 This trial has been used extensively in mock jury research and has been judged by experienced attorneys and trial judges to be a typical felony trial. See R. Hastie, S. Penrod & N. Pennington, Inside the Jury 47-50 (1983); Cowan, Thompson & Ellsworth, The Effects of Death Qualification on Jurors' Predisposition to Convict and on the Quality of Deliberation, 8 Law & Hum. Behav. 53, 63-64 (1984).
Figure 1. The Story Model for juror decision making.

(Figure 1 top left): (1) case-specific information acquired during the trial (e.g., statements made by witnesses about past events relevant to the decision); (2) knowledge about events similar in content to those that are the topic of dispute (e.g., knowledge about a similar crime in the juror's community); and (3) generic expectations about what makes a complete story (e.g., knowledge that human actions are usually motivated by goals). This constructive mental activity results in one or more interpretations of the evidence that have a narrative story form (Figure 1, top right). One of these interpretations (stories) will
be accepted by the juror as the best explanation of the evidence. The story that is accepted is the one that provides the greatest coverage of the evidence and is the most coherent, as determined by the particular juror.

1. Active Story Construction

When we hypothesize that jurors impose a narrative organization on evidence, we mean that jurors engage in an active, constructive, comprehension process in which evidence is organized, elaborated, and interpreted by them during the course of the trial. In part, this activity occurs because comprehension is inherently a constructive process for even the simplest discourse.11 This is especially true in the context of legal trials in which characteristics of the trial evidence make it unwieldy. First, there is a large amount of evidence, often presented over a duration of several days. Second, evidence presentation typically appears in a disconnected question and answer format; different witnesses testify to different pieces of the chain of events, usually not in temporal or causal order. Witnesses also are typically not allowed to speculate on necessary connecting events such as why certain actions were carried out, or a person’s emotional reaction to a certain event.

According to the theory, stories are constructed by reasoning from world knowledge and from evidence. Some potential story elements are accepted as true directly on the basis of their appearance as evidence from one or more credible sources; they are reasonably well established as facts. Which of these events will appear relevant depends upon the interpretation assigned to each fact with respect to its causal relatedness to other events. The inclusion in the story of other

---


This is the dominant view in cognitive psychology today. To illustrate, suppose a listener is told a simple narrative: “Billy went to Johnny’s birthday party. When all the children were there, Johnny opened his presents. Later, they sang Happy Birthday and Johnny blew out the candles.” Many listeners will infer spontaneously, and most will agree when asked, that there was a cake at the birthday party. Yet, no cake is mentioned in the sentences above; indeed, it is not certain that there was a cake. The cake is inferred because we share knowledge about birthday party traditions and about the physical world (the candles had to be on something). Another illustration comes with the comprehension of the sentence, “The policeman held up his hand and stopped the car.” Most of us understand this sentence in the cultural context of the policeman’s authority, shared signals, a driver watching the policeman but controlling the car, et cetera. Indeed, this is a sentence that would be puzzling to a person from a different culture.
evidence, inferred events, and causal relations between them is the result of a wide variety of deductive and inductive reasoning procedures applied to the evidence and world knowledge.\footnote{Collins, \textit{Fragments of a Theory of Human Plausible Reasoning}, in \textit{PROCEEDINGS OF THE CONFERENCE ON THEORETICAL ISSUES IN NATURAL LANGUAGE PROCESSING} II 2, 2 (D. Waltz ed. 1978); Collins \& Michalski, \textit{The Logic of Plausible Reasoning: A Core Theory} 13 COGNITIVE SCI. 1, 6 (1989).}

Analyses of inference chains leading to story events reveal that intermediate conclusions are established by converging lines of reasoning which rely on deduction from world knowledge, analogies to experienced and hypothetical episodes, and reasoning by contradiction.\footnote{N. Pennington \& R. Hastie, \textit{Representation and Inference}, supra note 7.} A typical deduction from world knowledge in the “Johnson case” consists of the following premise (P1 - P3) and conclusion (C) structure:

\begin{enumerate}
\item A person who is big and known to be a troublemaker causes people to be afraid.
\item Caldwell was big.
\item Caldwell was known to be a troublemaker.
\item Johnson was afraid.
\end{enumerate}

In this example, the juror matches features of Caldwell from undisputed evidence (P2) and a previous inferential conclusion (P3) to world knowledge about the consequences of being confronted with such a person (P1) to infer that Johnson was afraid (C).\footnote{It is the certainty of the conclusion C as a function of the levels of certainty of P1, P2, and P3, and the strengths of the relationships between the premises and conclusion that probabilistic (and heuristic) theories of inference were designed to model. It is at this point that Bayesian or Fuzzy Set calculations could be incorporated into the Story Model to yield the level of certainty with which a juror believes in any particular proposition (and consequently in the ultimate decision proposition). However, because of the lack of empirical support for Bayesian calculations as a \textit{description} of human judgment under uncertainty, we have adopted a set of simple assumptions that we believe are closer to actual juror judgment processes and will allow us to perform calculations over a network of relationships. The main assumption is that at the time an inferential conclusion is being considered as a potential story event, it is either regarded as certainly true (and therefore as data; e.g., P2), or as uncertain (and therefore as an hypothesis; e.g., P3, C), or as rejected and therefore certainly untrue. The final level of acceptability of any given proposition is hypothesized to be a function of its role in the story and its relation to relevant world knowledge (we return to the subject of juror confidence in the section on certainty principles, \textit{infra} pp. 527-28).}

Confidence in the conclusion drawn from an inference is assessed by analogizing the reasoning to other experiences and by evaluating alternate conclusions that would contradict the initial conclusion. For example, suppose that the same juror who provided the premise-conclusion example just mentioned, continued with, “If someone like Caldwell came up to me in a bar and threatened me, I would be afraid.” Alternate reactions were also considered, such as, “I don’t
think Johnson was angry. If he had been angry, he would have gone right back to the bar. He didn’t go right back.” This alternative is rejected: “No, Johnson was afraid of Caldwell and he took his knife with him because he was afraid.”

Different jurors will construct different stories, and a central claim of the theory is that the story will determine the decision that a particular juror reaches. Because all jurors hear the same evidence and have the same general knowledge about the expected structure of stories, differences in story construction must arise from differences in world knowledge; that is, differences in experiences and beliefs about the social world. In contrast to the example inference above, another juror might believe that confrontations by bullies are challenges to manly pride and that, as a result, anger is a more likely response. This particular inference may be a keystone in an evolving interpretation of the evidence that is completely different from that of the previous juror.

2. The Structure of Stories

Stories involve human action sequences connected by relationships of physical causality and intentional causality between events. In its loosest form, a story could be described as a “causal chain” of events in which events are connected by causal relationships of necessity and sufficiency. However, psychological research on discourse comprehension suggests that story causal chains have additional higher order structure both when considering the discourse itself and when considering the listener or reader’s “mental representations” of the discourse. Stories appear to be organized into units that are often called episodes. We show an abstract episode schema in Figure 2 that depicts a typical configuration of events in an episode; an episode should contain events which fulfill particular roles and are connected by certain types of causal relationships. In stories and episodes, events considered to be initiating events cause characters to have psychological responses and to form goals that motivate subsequent actions which cause certain consequences and accompanying states. An

---

example of an episode in the Johnson case is the following sequence: Johnson and Caldwell are in Gleason’s bar. Caldwell’s girl friend, Sandra Lee, goes up to Johnson and asks him for a ride to the race track the next day (initiating events). Caldwell becomes angry (internal response), pulls his razor, and threatens Johnson (actions; note that the goal is missing). Johnson backs off (consequence).

Figure 2. An abstract episode schema.

Stories may have further structure by virtue of the fact that each component of an episode may itself be an episode. For example, the entire episode above (characterized as Caldwell threatens Johnson) is the initiating event in one version of the Johnson story. In this version, the afternoon “threat” episode causes Johnson to be angry and want to pay Caldwell back. Thus, a story may be thought of as a hierarchy of embedded episodes. The highest level episode charac-

---

17 Rumelhart, supra note 16, at 277; Trabasso & van den Broek, supra note 15, at 627.
terizes the most important features of “what happened.” Components of the highest level episode are elaborated in terms of more detailed event sequences in which causal and intentional relations among subordinate story events are represented.

The structure of stories, according to our theory, plays an important role in the juror’s comprehension and decision-making processes. The story constructed by the juror will consist of some subset of the events and causal relationships referred to in the presentation of evidence, and additional events and causal relationships inferred by the juror. Some of these inferences may be suggested by the attorneys and some may be constructed solely by the juror. Whatever their source, the inferences will serve to fill out the episode structure of the story. Thus, expectations about the kinds of information necessary to make a story tell the juror when important pieces of the explanation structure are missing and when inferences must be made. Knowledge about the structure of stories allows the juror to form an opinion concerning the completeness of the evidence, or the extent to which a story has all its parts. In addition, the structure of episodes in a story corresponds to the structure of our knowledge about human action sequences in the world; story construction is a general comprehension strategy for understanding human action. Thus, the juror can easily compare the structure that is being imposed on the evidence to prior knowledge that is already encoded. Finally, the hierarchical episode and causal structure of the story provides an “automatic” index of the importance of different pieces of evidence. In the example above, the details of the embedded “threat” episode are subordinate in importance to the details of the top-level episode that reveal what Johnson did in order to pay Caldwell back. However, this indexing of importance is something that emerges from the structure of the story.

3. Certainty Principles

More than one story may be constructed by the juror. However, one story will usually be accepted as the “best” story. Furthermore, the juror will have a level of confidence in that “best” story which may be quite high or quite low. The principles that determine acceptability of a story and the resulting level of confidence in the story, are called certainty principles. According to our theory, two certainty principles govern acceptance: coverage and coherence. An additional certainty principle, uniqueness, will contribute to confidence.

A story’s coverage of the evidence refers to the extent to which

---

the story accounts for evidence presented at trial. Our principle states that the greater the story's coverage, the more acceptable the story as an explanation of the evidence, and the greater confidence the juror will have in the story as an explanation, if accepted. An explanation that leaves much of evidence unaccounted for is likely to have a lower level of acceptability as the correct explanation. Poor coverage should lower the overall confidence in a story, and consequently lower confidence in the decision.

A story's coherence also affects its acceptability and the level of confidence it induces given that the story is accepted. However, coherence is a concept in our theory that has three components: consistency, plausibility, and completeness. A story is consistent to the extent that it does not contain internal contradictions either with evidence believed to be true or with other parts of the explanation. A story is plausible to the extent that it corresponds to the decision maker's knowledge about what typically happens in the world and does not contradict that knowledge. A story is complete when the expected structure of the story “has all of its parts” (according to the rules of the episodic structure; see Figure 2 and discussion above). Missing information, or lack of plausible inferences about one or more major components of the story structure, will decrease confidence in the explanation. Thus, the coherence of the explanation reflects the consistency of the explanation with itself and with world knowledge, and the extent to which parts of the explanation can be inferred or assembled. These three ingredients of coherence—consistency, plausibility, and completeness—may be fulfilled to a greater or lesser degree and the values of the three components will combine to yield the overall level of coherence of a story.\footnote{Our concept of the coherence of a particular story is closely related to the idea of “persuasiveness” of the evidence.}

Finally, if more than one story is judged to be coherent, then the stories will lack uniqueness, which contributes to confidence in a story and in a decision. If there are multiple coherent explanations for the available evidence, belief in any one of them over the others will be lessened.\footnote{Einhorn & Hogarth, \textit{Judging Probable Cause}, 99 \textit{Psychological Bull.} 3, 15 (1986); Van Wallendael, \textit{The Quest for Limits on Noncomplementarity in Opinion Revision}, 43 \textit{Organizational Behav. \& Hum. Decision Processes} 385, 392 (1989).} If there is one coherent story, this story will be accepted as the explanation of the evidence and will be instrumental in reaching a decision.\footnote{These principles have been elaborated and formalized. N. Pennington, P. Messamer \& R. Nicolich, Explanatory Coherence in Legal Decision Making (unpublished manuscript 1991) (copy on file with author).}
4. Summary

Meaning is assigned to trial evidence through the incorporation of that evidence into one or more plausible stories which describe "what happened" during events testified to at the trial. General knowledge about the structure of human purposive action sequences and stories, characterized as an episode schema, serves to organize events according to the causal and intentional relations among them as perceived by the juror. Specific world knowledge about events similar to those in dispute will determine which particular interpretation is constructed or accepted. The level of acceptance will be determined by the coverage, coherence, and uniqueness of the "best" story.

B. Learning Verdict Definitions

The second processing stage in the juror's decision, according to the Story Model, involves the comprehension and learning of the decision alternatives, which in criminal trials are the definitions of the verdicts (e.g., first-degree murder, second-degree murder, et cetera). Most of the information for this processing stage is given to jurors at the end of the trial in the judge's substantive instructions on the law, although jurors may also have prior ideas concerning the meanings of the verdict categories (see Figure 1, bottom).

The verdict definitions in the judge's instructions are usually abstract and often couched in unfamiliar language. A crime is named and then abstract features are presented that define the crime. Features typically describe requirements of identity, mental state, circumstances, and actions that constitute the crime. For example, a judge's definition of first-degree murder presented as a feature list is shown in Figure 3.

We hypothesize that the juror's mental representation of this information also takes the form of a category label with a list of features. In all respects, this is a difficult one-trial learning task. If the juror has no prior knowledge of the legal categories, then learning of the abstract information is extremely difficult. In the case where prior knowledge is available, it is equally likely to interfere with accurate understanding as it is to help, because jurors' prior exposures to concepts such as first-degree murder, manslaughter, armed robbery, et cetera, are often (mis-) informed by television episodes and other media presentations.

FIRST DEGREE MURDER

IDENTITY: -RIGHT PERSON

MENTAL STATE: -INTENT TO KILL
-PURPOSE FORMED

CIRCUMSTANCES: -INSUFFICIENT PROVOCATION
-INTERVAL BETWEEN RESOLUTION AND KILLING

ACTIONS: -UNLAWFUL KILLING
-KILLING IN PURSUANCE OF RESOLUTION

Figure 3. Example verdict category represented as a feature list.

C. Making a Decision

The third processing stage in our hypothesis regarding the juror’s decision making involves matching the accepted story with each of the verdict definitions. In cognitive processing terms, this is a classification process in which the best match between the accepted story’s features and verdict category features is determined (see Figure 1, middle).

Because verdict categories are unfamiliar concepts, the classification of a story into an appropriate verdict category is likely to be a deliberate process. For example, a juror may have to reason about whether a circumstance in the story, such as “pinned against a wall,” constitutes a good match to a required circumstance, such as “unable to escape,” for a verdict of not guilty by reason of self-defense.

Although difficult, the classification process is aided by relatively direct relations between attributes of a verdict category (crime elements) and components of the episode schema (see Figure 4). The law has evolved so that the main attributes of the decision categories suggested by legal experts—identity, mental state, circumstances,

---

23 Id. at 1-4.
and actions—correspond closely to the central features of human action sequences represented as episodes—initiating events, goals, actions, and states. This is not a coincidence; rather, it is a reflection of the fact that both stories and crimes are culturally determined generic descriptions of human action sequences.

The story classification stage also involves the application of the judge’s procedural instructions on the presumption of innocence and the standard of proof. If the best fit is above a particular threshold requirement, then the verdict category that matches the story is selected. If some or none of the verdict attributes for a given verdict category are satisfied “beyond a reasonable doubt” by the events in the accepted story, then the juror should presume innocence and return a default verdict of not guilty.24

1. Certainty Principle

A further assessment of confidence occurs in the story classification stage. An evaluation of goodness-of-fit between the story and the best-fitting verdict category is based on the extent to which the story includes instantiations of elements of the category. The more missing element matches between the components of the episode schema and the attributes of the verdict category (see Figure 4), the lower the juror’s confidence in the verdict. As we speculated above, if the goodness-of-fit is not sufficient, then a default decision will be made.

D. Temporal Relations Between Stages

The processing stages have been presented as though a story has been constructed, then the verdicts are represented, and then a decision is reached. A fundamental claim of our theory is that the explanation structure is created a priori and causes the decision. The explanation structure is not a structure that is developed as a post hoc justification of the decision. This does not preclude a version of the theory in which there is cycling through the decision phases more than once; in such a case there could be an influence of the tentative decision (initial verdict classification) on an elaboration of the explanation. For example, story construction probably does not stop abruptly at the conclusion of the presentation of evidence. Previous

---

24 We have no empirical evidence on this point. We are basing this on the assumption that jurors either: (a) construct a single “best” story, rejecting other directions as they go along, or (b) construct multiple stories and pick the “best.” In either case, we allow for the possibility that the best story is not good enough or does not have a good fit to any verdict option and, therefore, a default verdict would have to be available. Ronald Allen has suggested another possibility for criminal cases—that the juror will search for any plausible story consistent with innocence. Letter from Ronald Allen to Reid Hastie and Nancy Pennington (Apr. 17, 1991).
Figure 4. The main elements of a story (Episode Schema) map onto the defining attributes of a verdict definition (Verdict Category Attributes).
research suggests that jurors' judgments involve much sifting and weighing of evidence as well as reflection on the meaning of the verdict categories after the major courtroom events of the trial are concluded. In addition, story meanings are not static structures. Although many causal and intentional inferences are made during the initial comprehension and encoding of events, causal information processing may not be completed during comprehension. Rather, causal reasoning begins then and continues with subsequent attributional inferences influenced by the initial comprehension process. Examining the extent to which processing stages interact is a long-term goal of our research.

II. EMPIRICAL EVIDENCE FOR THE STORY MODEL

The basic claim of the Story Model is that story construction enables critical interpretive processing and organization of the evidence so that evidence can be meaningfully evaluated against multiple verdict judgment dimensions. The Story Model provides a psychological account for the assignment of relevance to presented and inferred information. Precise claims are made concerning the representational form of the evidence, and a mediating role is claimed for stories in subsequent decisions and confidence in those decisions. Uncertainty in the decision is centered in assessments of the coverage, coherence, and uniqueness of the story, and on the goodness-of-fit of the story with reference to the verdict categories. Detailed summaries of empirical studies of the claims of the Story Model are provided in other reports. In this section, we summarize the empirical support for the theory.

A. Interview Study

Our initial research on the Story Model was designed to elicit data that would provide a snapshot of the juror's mental representations of evidence and of verdict categories at one point in time. Three questions were the focus of the empirical analysis: (1) Do the mental representations of evidence show a story structure?; (2) Do the mental

25 R. Hastie, S. Penrod & N. Pennington, supra note 10; N. Pennington, Causal Reasoning, supra note 7.

26 W. Kintsch, supra note 11, at 170-71.

27 Pennington & Hastie, Evidence Evaluation, supra note 7; Pennington & Hastie, Memory Structure, supra note 7; Pennington & Hastie, Explaining the Evidence, supra note 7; N. Pennington, Causal Reasoning, supra note 7; N. Pennington & R. Hastie, Story Structure, supra note 7.

28 Pennington & Hastie, Evidence Evaluation, supra note 7; N. Pennington, Causal Reasoning, supra note 7.
representations of verdicts show a category structure?; (3) Are there systematic relationships between an individual juror's verdict decision and that same juror's evidence representation, verdict representation, or classification procedures?

The first study was based on a correlational logic, using an interview to provide structural descriptions of mental representations and then determining whether or not the structures covaried systematically with verdict decisions. According to the Story Model, if story construction is a central determinant of verdict decisions, then we should find that variability in verdict decisions correlates with variability in story structures and is not related to verdict category representations or classification procedures.

Twenty-six adult subjects were sampled from volunteers in the Massachusetts Superior Court jury pool and shown a realistic filmed reenactment of a murder trial (the Commonwealth v. Johnson case described earlier). Subjects were instructed that an actual jury had decided the case and their task was to "be one of the jurors" and to try to reach a decision on the verdict. In the trial, the jurors chose from among four verdict alternatives in reaching a decision. The defendant Frank Johnson could be judged to be not guilty, guilty of manslaughter, guilty of second-degree murder, or guilty of first-degree murder. A sub-sample of sixteen of the twenty-six subjects was chosen for extensive analysis so that a range of verdicts was obtained.²⁹ The verdict distribution for the sixteen subjects was: five guilty of first-degree murder; four guilty of second-degree murder; four guilty of manslaughter; three not guilty (self-defense). The major source of data was a verbal protocol obtained in an interview with each experimental juror asking each to think aloud while making a decision and to respond to questions about the evidence and about the judge's instructions to the jurors.

Verbal protocols from each juror were analyzed by coding all assertions about events and relationships between events that were claimed to have occurred or not to have occurred within the context of the events referred to in testimony. Many of these assertions referred to events and relationships actually mentioned in testimony and many referred to events and relationships inferred by the juror. This coding was transformed into a directed graph designating interrelations between events.³⁰ A graph structure was created for each

²⁹ See Pennington & Hastie, Evidence Evaluation, supra note 7; N. Pennington, Causal Reasoning, supra note 7.

³⁰ Goodman & Hedetniemi, A Descriptive Introduction to Graph Theory and Some of its Applications: Formal Methods, in COMPUTERS IN LANGUAGE RESEARCH 19 (W. Sedelow & S.
subject in which the nodes represented event codes and the links represented the asserted connections between events. This structure captured part of each subject’s conceptual representation of the evidence as indicated by the protocol events mentioned and assertions regarding relations between events. To facilitate economy of presentation and to permit comparisons with other experiments, only the results from the two extreme verdict groups—first-degree murder and not guilty—will be summarized in the present report. Analyses of the data for subjects from all four verdict groups are in agreement with this summary.

Before reviewing evidence that these graphs had a story structure, we should ask what range of plausible structures we might expect. First, evidence could be stored in memory in an unembellished form, as it was presented at trial—in a disconnected question-and-answer sequence, organized by witness and interconnected largely by referential coherence. This is plausible because we know that when judgments are made on-line, memory for evidence is unrelated to the judgment. Second, the evidence could be conceptualized in terms of the structure of the legal argument, as analyzed by legal scholars and other theorists. In this conception, evidence is structured in terms of arguments for and against guilt; or, for and against a re-

32 Pennington & Hastie, Evidence Evaluation, supra note 7; N. Pennington, Causal Reasoning, supra note 7.
33 Making a judgment “on-line” means incorporating the value of a piece of evidence into a judgment as soon as it is encountered. In a legal trial context, this means that when a witness testifies, “Johnson was carrying a knife,” the juror immediately increases his or her belief in guilt. If the witness says, “It was a fishing knife,” the juror immediately decreases belief in guilt. Story construction is not an “on-line” decision strategy, but rather a “memory-based” strategy because evidence is organized, elaborated, and interpreted in memory before entering into a judgment.
34 For an example of such structure, see Pennington & Hastie, supra note 6, Fig. 3, at 256.
quired element of guilt with respect to a particular charge. A third possibility is that the important evidence revolves around the characterizations of the defendant and victim. In this case, structures emerging from protocols would show character sketches connected to verdicts. A final possibility is our theory—that the juror organizes the evidence into a story that emphasizes the causal and intentional relations among evidence items.

Our first major conclusion from the interview study was that the mental representations of evidence derived from the interview protocols showed story structures and not other plausible structures. There were several features of the conceptual graph structures that support our claim that these structures had story form and not one of the other plausible forms. First, 85% of all the events referred to in the protocols were causally linked. Thus, subjects were primarily making assertions like, “Johnson was angry so he decided to kill him” (here anger initiates the goal to kill), rather than assertions like, “Johnson was a violent man. That makes me think he intended to kill him” (the fact that Johnson was violent leads to an inference of intention to kill). This is strong evidence that subjects were telling stories, and not constructing arguments.\(^{37}\) Second, only 55% of the protocol references were to events that were actually included in testimony. The remaining 45% were references to inferred events—actions, mental states, and goals—that “filled in” the stories in episode configurations. This data contradicts the image of the juror as a “tape recorder” with a list of trial evidence in memory. Experimental jurors did make character inferences (5.4% of story content), but these were integrated into the story structures as reasons for certain behaviors. For example, an inference that Caldwell was a violent man might be given as a reason that Caldwell pulled a razor when provocation was slight, or as a reason that Johnson was afraid. Finally, the conceptual graphs could be represented as hierarchies of embedded episodes when rules were applied to identify explicit goals linked to actions leading to final consequences.\(^{38}\) Examples of these structures are illustrated in Figure 5.

The second major conclusion from the interview study was that story structures differed systematically for jurors choosing different verdicts. In order to analyze this phenomenon, a measure of shared


\(^{38}\) See N. Pennington, Causal Reasoning, supra note 7.
features\textsuperscript{39} was used to develop a central story for each verdict group which we call verdict stories. For example, the central story for the jurors choosing first-degree murder is the first-degree murder verdict story. A network was assembled containing only those event codes and links shared in common by 80\% of the members of the verdict group. An episode structure was imposed on the causal chains by applying rules to identify explicit goals linked to actions leading to the final consequence.\textsuperscript{40} Verdict stories for first-degree murder and not guilty verdict groups are shown in Figures 5A and 5B.

The gist of the first-degree murder verdict story (Figure 5A) is that an argument and threat by Caldwell (the victim) so enraged Johnson (the defendant) that Johnson decided to kill Caldwell. Johnson got his knife, found Caldwell, got into a fight, and stabbed him to death. In contrast, the gist of the not guilty story (Figure 5B) is that Caldwell started a fight with Johnson and threatened him with a razor. Johnson used a knife in order to protect himself and Caldwell ran into the knife.

The episode structures of the two stories map neatly onto their respective verdict category attributes. For example, in the not guilty verdict story (Figure 5B), there are three episodes, two of which are embedded. The main episode is the fight, and the initiating events are all of Caldwell’s actions during the fight. The afternoon episode serves to fortify not guilty subjects’ conclusions about Johnson’s psychological state, leading first to a goal to show the knife and then to actively protect himself. The not guilty story shows the knife going into Caldwell as a consequence rather than as a goal-directed action. These features correspond to the verdict features of not guilty by reason of self-defense: under immediate attack, unable to escape, intent to defend, and reasonable retaliation. First-degree murder requires premeditation; that is, a resolution formed to kill, an interval of time, and a killing in accordance with the resolution. The subjects’ emphasis on the initiation of an intent-to-kill goal is expressed through the elaborated afternoon events (Figure 5A). Thus, being hit is not an initiating event, but part of a sequence of acts that follow from behavior directed by a goal to kill.

Verdict representations were coded and compared across different verdict groups. The majority of references to verdicts during the portion of the interview when subjects spoke aloud were to particular

\textsuperscript{39} See Tversky, Features of Similarity, 84 Psychological Rev. 327, 329 (1977); see generally Tversky & Gati, Studies of Similarity, in Cognition and Categorization 79 (E. Rosch & B. Lloyd ed. 1978).

\textsuperscript{40} N. Pennington, Causal Reasoning, supra note 7.
Figure 5A. First-degree murder verdict story (central story for jurors choosing first-degree murder).

Figure 5B. Not guilty verdict story (central story for jurors choosing not guilty by reason of self defense).
category features such as “premeditation” and “malice.” When asked to relate what the verdicts meant, jurors provided lists of features—although they were usually far from complete. Jurors also elaborated their category definitions in some cases by constructing mini-stories to illustrate. An example is, “First-degree murder is premeditated. There is a plan. That would be if he had gone back to the bar looking for him.”

There is considerable variation among jurors in the accuracy and completeness of their representations of the verdict category information. If these variations are systematically related to the jurors’ verdict choices, then the hypothesis must be retained that differences occurring in the verdict learning stage determine the verdict choice. For example, if jurors favoring first-degree murder verdicts were also jurors who did not remember the judge’s instruction concerning premeditation, it would be plausible that the memory failure affected the verdict choice. To the contrary, analyses of the answers to questions about verdict category definitions showed that juror verdict choice was not related to memory for verdict-element relations. Other analyses showed no content differences by verdict choice. Thus, variance in performance in the verdict learning stage of the juror decision does not appear to determine the juror’s decision.

The interview study did not obtain information concerning the jurors’ notions of presumption of innocence or the “beyond reasonable doubt” standard. However, other research sampling jurors from the same population of jurors and using the same stimulus trial did obtain direct ratings of subjects’ estimated values for the “beyond reasonable doubt” threshold and the “presumption of innocence” principle. Analyses conducted on these ratings did not indicate significant differences among the subjects when they were classified by verdict choice. The implication for the present work is that variation in performance in the story classification stage is not associated with variations in verdict choice.

In this research, two key results were established that were necessary conditions for pursuit of the Story Model as a viable theory of decision making in the juror context. First, the evidence structures constructed by jurors had story structure (not other plausible structures) and verdict structures looked like feature lists. Second, jurors who chose different verdicts had constructed different stories. Thus, decisions covaried with story structures, but not with verdict representations or story classification processes.

---

41 Pennington & Hastie, Evidence Evaluation, supra note 7.
42 R. Hastie, S. Penrod & N. Pennington, supra note 10.
The interview study served its purpose as our initial investigation and played a role in our theory building. Moreover, it provides a strong empirical foundation for the Story Model using a realistic stimulus trial and a range of adult citizen jurors. However, the interview methodology itself may have created a demand for stories as conversational forms or as justifications. Subsequent experiments addressed this issue and tested additional claims of the theory.

B. Predicting Importance Ratings and Memory for Evidence

A second empirical study was conducted to test the conclusions of the interview study using conventional laboratory research methods with college student subjects. In this study, subjects' responses to sentences presented in a recognition memory task were used to identify subjects' post-decision representations. The major motivation for this study was to test whether stories were constructed spontaneously in the course of the juror's performance. A second goal of the study was to test our claim that the position of an evidence item in its verdict story would predict the importance rating for that item according to verdict choice.

Stimulus materials were constructed based on the content analysis of interviews from the first study. This yielded a 119 sentence written version of the Commonwealth v. Johnson case. The written case was carefully constructed so that its constituent sentences included propositions from each verdict story that were not also present in any other verdict story. For example, the proposition “Johnson stabbed Caldwell” was a part of the first-degree murder verdict story, but not included in the not guilty verdict story. In addition, sentences were prepared for the recognition test that were not presented in the stimulus case, but which were parts of the verdict stories identified in the interview study; that is, they were frequently inferred by the jurors choosing a particular verdict. Thus, a recognition test could be constructed with old (true) target sentences from each story that had been presented as evidence, and with new (false) lures from each story that had not been presented as evidence (but

---

43 Pennington & Hastie, Memory Structure, supra note 7.

44 This is an empirical observation. This particular statement could have, in principle, been in any of the verdict stories. We described earlier how verdict stories were determined empirically. The graph of events and links for each juror choosing the same verdict were compared. Those events and links in the stories of 80% of the jurors in the verdict group were retained as part of the verdict story. Some of the constituent events and links were evidence items and some were inferences shared by jurors in the verdict group. A particular verdict story would contain only part of the total body of evidence and only part of the total body of inferences that various jurors drew.
were frequent inferences.\textsuperscript{45}

Subjects "recognized" sentences from the story associated with their verdict as having been presented as trial evidence with a higher probability than sentences from stories associated with opposing (rejected) verdicts, for both old and new items. That is, they were more likely to correctly recognize as evidence those evidence items in the verdict story corresponding to their own verdict choices, and they were more likely to falsely recognize as evidence those inferences in the verdict story corresponding to their decisions.

Subjects also rated the importance of trial evidence items. These ratings were strongly related to the causal role of the item in the story associated with a subject's verdict. These results corroborated the conclusions about story structure and story-verdict relationships from the initial study. Furthermore, they implied that story representations were constructed spontaneously as part of the natural decision process, and not artificially elicited by the interview task used in the first study.

Even though we demonstrated with this experiment that causal explanations are constructed spontaneously in the course of decision making, we have still not demonstrated that the story constructed by the juror is a true mediator of the decision. It could still be the case that the juror makes a decision and then constructs a story as a post hoc justification. The next experiments address this question.

\textbf{C. Predicting Decisions and Confidence in Decisions}

Two experiments were conducted to test our claim that stories cause decisions. We reasoned that if we could manipulate the ease of constructing a particular story and thereby influence the likelihood of the corresponding decision, then this would be strong evidence for our claim of causal mediation. In both studies, we manipulated the ease of constructing a particular story by varying the presentation order (but not the content) of the evidence.

In our third study,\textsuperscript{46} using the abbreviated Commonwealth v. Johnson stimulus trial, we varied presentation order to influence the

\textsuperscript{45} Examples of each type of recognition memory test item are as follows:

\begin{enumerate}
\item[I.] Old items (presented as evidence) and empirically determined to be a part of:
\begin{enumerate}
\item[A.] Not Guilty Verdict Story: Johnson held his knife out in front of himself.
\item[B.] Murder Verdict Story: Johnson stabbed Caldwell in the chest.
\end{enumerate}
\item[II.] New items (not presented as evidence) but empirically determined to be a frequent inference in:
\begin{enumerate}
\item[A.] Not Guilty Verdict Story: Johnson was trying to protect himself.
\item[B.] Murder Verdict Story: Johnson was looking for Caldwell.
\end{enumerate}
\end{enumerate}

\textsuperscript{46} Pennington & Hastie, \textit{Memory Structure}, supra note 7.
ease with which a prosecution (guilty of murder) or defense (not guilty by reason of self-defense) story could be constructed. Stories were considered easy to construct when the evidence was ordered in a temporal and causal sequence that matched the occurrence of the original events (story order). Stories were considered difficult to construct when the presentation order did not match the sequence of the original events. We based the non-story order on the sequence of evidence as conveyed by witnesses in the original trial (witness order). The logic of the experiment was summarized in our hypothesis that (manipulated) ease of story construction would influence verdict decisions; easy to construct stories would result in more decisions in favor of the corresponding verdicts.

One hundred and thirty college student mock jurors listened to a tape recording of a one hundred-item version of the Commonwealth v. Johnson evidence (50 prosecution statements and 50 defense statements), followed by a judge’s charge to choose between a guilty of murder verdict and a not guilty verdict. The 50 prosecution statements and the first-degree murder story were presented either in story order or witness order. Similarly, the defense statements were presented in one of the two orders creating a four-cell factorial design. In all four order conditions the prosecution evidence preceded the defense evidence as per standard legal procedure. After listening to the tape-recorded trial materials, the mock jurors completed a questionnaire indicating their verdict, their confidence in the verdict, and their perceptions of the strengths of the prosecution and defense cases.

As predicted, mock jurors were most likely to convict the defendant when the prosecution evidence was presented in story order and the defense evidence was presented in witness order (78% chose guilty), and they were least likely to convict when the prosecution evidence was in witness order and the defense evidence was in story order (31% chose guilty; see Table 1). Thus, story coherence, as determined by presentation order of evidence, affects verdict decisions in a dramatic way.47

Analyses conducted on the strength ratings of the defense and prosecution cases reveal that these ratings were influenced by presentation order, with story order evidence rated stronger than witness order. Furthermore, the perceived strength of one side of the case

47 See V. Smith, The Psychological and Legal Implications of Pre-Trial Instructions in the Law, (Ph.D. dissertation, Stanford University 1987) (abstract found in 48 DISSERTATION ABSTRACTS INTERNATIONAL 3451B (1988)) (for a replication of this study with different materials comparing story order with an organization of evidence by legal issue rather than by witness order).
Table 1

*Percentage of Subjects Choosing a Verdict of Guilty of Murder by Prosecution and Defense Order Conditions*

<table>
<thead>
<tr>
<th>Defense Evidence</th>
<th>Story Order</th>
<th>Witness Order</th>
<th>Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Story Order</td>
<td>59%</td>
<td>78%</td>
<td>69%</td>
</tr>
<tr>
<td>Witness Order</td>
<td>31%</td>
<td>63%</td>
<td>47%</td>
</tr>
<tr>
<td>Means</td>
<td>45%</td>
<td>70%</td>
<td></td>
</tr>
</tbody>
</table>

depended on the order of evidence for both that side and the other side of the case. This finding supports our claim that the uniqueness of the best-fitting story is one important basis for confidence in the decision. We also examined the verdict confidence ratings and found that, regardless of verdict chosen, jurors who heard both sides of the case in story order were more confident than jurors who heard one or neither side in story order. This result reinforces our conclusion that alternate story strength is also important, although the empirical finding was not predicted.

It should be noted that this experiment was a laboratory experiment designed to test our hypotheses about the presence or absence of certain mental processes and their relationship to the decisions made. The study was not designed to estimate the size of order effects in real trials. In real trials, there are many devices that assist the juror in story construction: opening and closing statements, redundancy in presentation of information, a rich visual environment, and so forth. In this particular experiment, we stripped those enhancements away to reveal the effect of making a story very difficult to construct. In order to estimate the size of order effects in actual trials, this type of experiment would need to be repeated with more realistic stimulus materials.

Our fourth experiment focussed again on the effects of variations in evidence presentation order, allowing us to further examine the relationship between evidence organization, memory organization, recall memory, and judgments.48 Using case materials developed by Devine and Ostrom,49 evidence was presented either in story order or

legal issue order. Two cases were used that varied in whether the preponderance of evidence favored guilt or innocence.

Materials were presented to 414 college student subjects in written form. The two evidence organizations resulted in different memory organizations of evidence, as measured by an analysis of clustering in free recall (using an “adjusted ratio of clustering”). Subjects who heard the evidence organized by story showed high story clustering and low issue clustering in free recall; subjects who heard the evidence organized by legal issue showed low story clustering and high issue clustering. However, total amount recalled was not different for the two evidence organization conditions.

The results replicated and extended our previous results. When stories were easily constructed, and therefore represented more coherently in memory, more verdicts were chosen in the expected direction and subjects rated their confidence as higher. Moreover, this effect was obtained in the absence of effects on the overall amount of recall, ruling out the hypothesis that manipulating order merely manipulates memorability of the evidence.

We had predicted and observed sizeable effects of story coherence on verdict choice and confidence that were consistent with the Story Model. Furthermore, the effects appeared in two very different sets of case materials. Alternate algebraic models for juror judgment, derived from the Bayesian approach and primacy-recency principles of information integration, could not account for the particular pattern of order effects we obtained. In our next experiments, we utilized a new set of experimental materials and tested predictions from the Story Model, a Bayesian model, and a Sequential Updating Model.

D. Comparisons With Other Models

Many psychological and legal analyses of the juror’s task postulate that the decision depends on the estimation and combination of probabilities. A variety of important legal concepts make reference to the probabilistic nature of evidence at trial. For example, relevant evidence is defined as evidence that has a tendency to make the exist-

---

This is computed by counting the number of times two statements in the same “story” are recalled together and expressing this number as a proportion of total recall, with both numerator and denominator adjusted for the expected number of story items that would occur together by chance if recall order was a randomly scrambled sequence of items. Similarly, an ARC would be computed for items recalled together that referred to the same legal “issue.”
51 Pennington & Hastie, supra note 6.
ence of a fact that is of consequence to the determination of the action more probable or less probable than it would be without the evidence.\textsuperscript{52}

There is no doubt that the juror’s task involves uncertainty, as do all complex decision tasks of the kind we are considering. However, treating the task as a probability assessment task assumes that the uncertainty assessments behave in ways that are consistent with the rules of mathematical probability theory. Within the mathematical (Pascalian) probability system there is a prescription for coherent probability revision in the light of evidence (Bayes’ rule). Under this prescription, probabilistic opinion revisions have three basic properties: the combining process is multiplicative, probabilities of alternate hypotheses must sum to one, and a hypothesis that is held at any time with a probability of zero cannot be revived.\textsuperscript{53} It is generally recognized that the Bayesian system is an invalid description of human behavior under most conditions.\textsuperscript{54}

An alternative probability model has been proposed by Cohen,\textsuperscript{55} the inductive probability system, in which probabilities have only ordinal properties. Negation is not complementary, zero probabilities correspond to “no reason to believe” and can therefore be revived with further evidence, and the opinion revision process is not multiplicative. Schum and Martin, in a recent test of the descriptive adequacy of both the Bayesian and inductive probability systems as theories for juror judgments stated, “we can be fairly conclusive in saying that our subjects did not typically respond in accordance with the canons of probabilistic inference in either the Baconian or Pascalian probability systems.”\textsuperscript{56} In general, features of human uncertainty assessment found across many tasks are inconsistent with the rules of one or more of the traditional probability calculi. For example, the subjective probabilities of complementary hypotheses have been found not to sum to one;\textsuperscript{57} if certainty about one hypothesis increases, certainty

\textsuperscript{52} FED. R. EVID. 401; Lempert, Modeling Relevance, 75 Mich. L. Rev. 1021 (1977).


\textsuperscript{54} Fischhoff & Lichtenstein, Don’t Attribute This to Reverend Bayes, 85 Psychological Bull. 239 (1978); Rappoport & Wallsten, Individual Decision Behavior, 23 Ann. Rev. Psychology 131 (1972); Slovic & Lichtenstein, Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment, 6 Organizational Behav. & Hum. Performance 649 (1971).

\textsuperscript{55} See L. Cohen, supra note 4.

\textsuperscript{56} D. Schum & A. Martin, supra note 53, at 77. See Einhorn & Hogarth, Ambiguity and Uncertainty in Probabilistic Inference, 92 Psychological Rev. 433, 458 (1985); Pennington & Hastie, Memory Structure, supra note 7, at 530.

\textsuperscript{57} Edwards, Subjective Probabilities Inferred From Decisions, 69 Psychological Rev.
about alternative hypotheses may remain constant, increase, or decrease. The subjective certainty attached to a conjunction of events is frequently overestimated relative to the optimal combination of the component uncertainties, indeed, the subjective certainty attached to a conjunction of events may be assessed to be greater than the certainty of one or more of the component events. Subjective certainty assessments may be too high under conditions where there is a high similarity between the pattern of evidence and a known standard, or when there is high internal consistency of the evidence even though the evidence is known or thought to be unreliable.

The alternative to modeling juror inference as a probabilistic opinion revision process is to consider that the weight of the evidence "accumulates" in some other manner. In this regard, the additive models are more consistent with the anomalies listed above than are the probability formulations.

In our final two experiments on the story construction process, we examined more closely the impact of story completeness on subjects' beliefs in the guilt of the defendant and its effect on evidence evaluation when subjects were asked to respond to the evidence at different levels of aggregation. We expected that more complete stories would produce more verdicts in the direction of the completed story. We also expected that there would be a greater effect of mediating story structures when evidence was evaluated globally at the end

109, 130 (1962); Einhorn & Hogarth, supra note 56, at 457; Robinson & Hastie, Revision of Beliefs When a Hypothesis is Eliminated From Consideration, 11 J. EXPERIMENTAL PSYCHOLOGY: HUM. PERCEPTION & PERFORMANCE 443, 450 (1985); D. Schum & A. Martin, supra note 53, at 47; Van Wallendae, supra note 20, at 390; Van Wallendae & Hastie, Tracing the footsteps of Sherlock Holmes: Cognitive representations of hypothesis testing, 18 MEMORY & COGNITION 240, 240 (1990).

58 Robinson & Hastie, supra note 57, at 450; D. Schum & A. Martin, supra note 53, at 47.

59 D. Schum & A. Martin, supra note 53, at 48.


62 Saks & Kidd, supra note 2, at 132-33; Schum & Martin, supra note 36, at 134-37; see generally Schum, DuCharme & DePitts, Research on Human Multistage Probabilistic Inference Processes, 10 ORGANIZATIONAL BEHAV. & HUM. PERFORMANCE 318 (1973); Tversky & Kahneman, Judgment under Uncertainty: Heuristics and Biases, 185 SCIENCE 1124 (1974).

63 For example, the "Information Integration Model" averaging rule and the "Sequential Weighting Model," are reviewed by Pennington & Hastie, supra note 6, at 257, 273.

64 See Pennington & Hastie, Explaining the Evidence, supra note 7.
of all of the evidence compared to judgments rendered after each item of evidence was presented. In addition, we compared these two decision modes, subjects’ *global* judgments\(^65\) (the normal decision mode for legal judgments), and their cumulative *item-by-item* judgments,\(^66\) with Bayesian, sequential updating and story models of aggregation.\(^67\)

We directly varied the ease of constructing particular stories by providing or withholding evidence that was of specific relevance to one possible story or another. These “evidence supplements” were designed to instantiate a component of either the defense or the prosecution story by strengthening causal links between certain pieces of evidence and/or weakening others. In this way, we expected to alter the interpretation of the evidence, thus leading to different decisions. The case materials and methods in these two experiments were based on work by Schum and Martin.\(^68\) Three evidence conditions were created for two of their cases involving an embezzlement and a burglary: a *convict* supplements version, an *acquit* supplements version, and the original materials from Schum and Martin (*basic* version).\(^69\) For the first experiment, following methods laid down by Schum and Martin,\(^70\) we had subjects respond to the case materials at three levels: a *global* assessment of the entire collection of evidence; *local* assessments of each block of evidence (essentially each witness’s testimony); and an *item-by-item* evaluation where the subject responded after each block of evidence indicating his or her current cumulative judgment. Because the supplements tied evidence together into a more (or less) coherent story, we expected that their effect would be greater when considered in the context of all the evidence (global judgment) than when their impact was incorporated into the judgment as the

---

\(^{65}\) A *global* judgment refers to the condition in which subjects read through the entire body of evidence and made a single evaluation of the likelihood of guilt at the end of that reading.

\(^{66}\) Cumulative *item-by-item* judgments refer to the condition in which subjects were asked to read a single block of evidence and then make a judgment of the likelihood of guilt, read the next block, and then make a new judgment (based on all evidence up to that point).

\(^{67}\) We compared these models to subjects’ actual judgments by including a third *local* judgment condition in which subjects were asked to rate the probative value of each evidence block independently. We then applied the three model combination rules to these local judgments and compared the model aggregation result to the subjects’ actual global and item-by-item judgments. The Bayes combination rule is well-known; each evidence block was considered to be independent. For a formal analysis of these stimuli, see D. Schum & A. Martin, *supra* note 53; Schum & Martin, *supra* note 36. The sequential updating model was an Anchor-and-Adjust model in which the current judgment (which is a summary of all previous judgments) was weighted .45 and the current evidence block was weighted .55. The model predicts large recency effects. The Story Model combination rule used equal weighing of evidence; that is, the probative evaluation of the item was its effective weight.

\(^{68}\) Schum & Martin, *supra* note 36.

\(^{69}\) D. Schum & A. Martin, *supra* note 53.

\(^{70}\) Schum & Martin, *supra* note 36.
evidence was heard (item-by-item judgment). This prediction was motivated by the assumption that when subjects are asked to make a single *global* judgment after reading the entire body of evidence, they are able to integrate evidence into a unitary summary structure before evaluation; that is, their judgment strategy will be “memory-based.”71 However, when subjects are asked to make a cumulative judgment after each evidence block, the subject is focussed on the adjustment or change in evaluation. This is likely to invoke an “on-line” strategy whereby subjects anchor on the current opinion and adjust for the new evidence confronting them.72 We also expected that neither the global nor the item-by-item judgments would be well fit by a Bayesian aggregation of the local evidence evaluations; that global judgments were more likely to have involved story construction; and that the item-by-item judgments would be better described by an anchor-and-adjust process.

During the two experiments,73 our prediction that the addition of story supplements would cause subjects to render stronger evaluations of evidence in the story direction was supported. That is, the convict version of the cases elicited greater odds in favor of guilt than the basic version and the basic version elicited greater odds in favor of guilt than the innocent version.

Next, we tested our assumptions about the strategies that subjects were using at different levels of aggregation. As predicted, the Bayesian model did not fare well as a description of subjects' global or item-by-item ratings in the experiment (also noted by Schum and Martin74). First, consistent with a hypothesis of “conservatism,” neither the final item-by-item nor the global ratings show the degree of influence of the evidence supplement manipulation that appears in the Bayesian calculation based on local evidence block ratings. Bayesian aggregates of the local judgments were about ten times stronger than the global evaluations and about fifteen times stronger than the item-by-item assessments. Thus the subject aggregates (global and item-by-item), consistent with previous research,75 are extremely conservative with respect to a Bayesian aggregation rule. Second, several

73 See Pennington & Hastie, *Explaining the Evidence, supra* note 7.
74 Schum & Martin, *supra* note 36.
specific qualitative characteristics of the item-by-item ratings, primarily in the form of non-complementary adjustments, contradict implications of the Bayesian rule.\textsuperscript{76} Third, direct comparisons of goodness-of-fit of a Bayesian updating model and an algebraic anchor-and-adjust model,\textsuperscript{77} applied to the item-by-item ratings, clearly favor the anchor-and-adjust model. The mean difference between item-by-item ratings and the anchor-and-adjust model over evidence blocks is not reliably different from zero for either stimulus case. In contrast, the best fitting algebraic description of the global ratings was neither the Bayesian nor the anchor-and-adjust model (differences between global ratings and anchor-and-adjust predictions were reliably different from zero). A configuration of weights consistent with the Story Model\textsuperscript{78} provided the best-fitting model for the global judgments.

We also predicted that story supplements would have greater impact on global judgments than on item-by-item judgments. This was supported by the fact that subjects' global assessments were stronger in force than the item-by-item final evaluation by a factor of about 1.5, and that the predicted interaction between evidence supplement treatments (convict versus acquit) and response modes (item-by-item versus global) on final judgments of guilt was obtained.

In sum, the essential results of the two studies were consistent with predictions from the Story Model and projections from closely related research. The Bayesian model did not provide an adequate description of human performance on either the final ratings of the global judgment task or the ultimate rating of the item-by-item response sequence. Nor did the Bayesian approach provide an acceptable description of item-by-item ratings across the course of evidence presentation. An anchor-and-adjust algebraic updating model did provide a satisfactory fit to the sequence of item-by-item judgments. The final item-by-item judgment was less polarized (as a function of the presence of acquit or convict evidence supplements) than the single global rating in all conditions. As our hypothesis predicted, anchor-and-adjust described the item-by-item judgment process, and story construction best described the global judgment.

E. Summary

The first study used an extensive interview to establish that intervening narrative structures were created by jurors in a realistic mock

\textsuperscript{76} See Schum & Martin, supra note 36; Pennington & Hastie, Memory Structure, supra note 7; Robinson & Hastie, supra note 57.

\textsuperscript{77} E.g., L. Lopes, supra note 72; Einhorn & Hogarth, supra note 56.

\textsuperscript{78} See Pennington & Hastie, Explaining the Evidence, supra note 7.
juror study; that these structures took the form of a story; that jurors who agreed on the verdict decision shared a common story; and that other traces of the decision process (e.g., estimates of standard of proof and knowledge of the verdict definitions) did not vary systematically with the decision. The second study using a recognition memory task reinforced the first study’s conclusions and added the finding that the story structures were created spontaneously, without the demands of communication with the experimenter in the interview situation.

The next two empirical studies provided substantial evidence that the story-like evidence summary is a key causal mediator of the verdict decision. In both studies, variations in the order of presentation of a fixed set of evidence had clear effects on verdict decisions. Furthermore, the order manipulations were selected to either facilitate or impede construction of conviction-prone or acquittal-prone stories yielding successful predictions of verdicts from evidence order via the Story Model. The overall pattern of verdict decisions, confidence ratings, and other collateral judgments also supported our hypothesis that completeness, coherence, and uniqueness of the best-fitting story would predict confidence in the correctness of the verdict.

The final empirical studies provided some comparisons of the Story Model to two traditional computation-oriented models, a Bayesian updating model and an algebraic anchor-and-adjust model. At the most general level, we hypothesized that the Bayesian formulation would not provide a satisfactory account of any of the human judgments; that the Story Model would describe global judgments based on all of the evidence; and that the anchor-and-adjust model would describe the sequence of judgments when subjects were prompted for cumulative ratings after each witness’s testimony. The general hypothesis and subsidiary hypotheses derived from the Story Model and the anchor-and-adjust model were confirmed.

III. Future Directions for the Story Model

The Story Model is not a “finished” theory. There are some aspects that have only been outlined and others that have not been addressed. We discuss five of these issues here: (1) issues of generalizability; (2) the need to elaborate process components of the model and possible temporal interactions between stages; (3) the need to further elaborate and test the determinants of confidence in decisions; (4) the role of stories as mediators of other influences in legal trials; and (5) the role of stories in group deliberation.
A. Generalizability

The issue of the generalizability of the Story Model has two parts. One part concerns questions of the extent to which the cognitive processes and mental structures proposed in the Story Model apply to a large range of legal cases. The second part concerns questions of generalizability of our theoretical principles to actual trial settings, and their implications. Concerning the first issue, our research on the Story Model has used case materials that could be considered especially conducive to story construction. For example, we have concentrated on criminal over civil cases, on cases not reducible to a single issue such as “who-did-it,” and on two-sided rather than one-sided cases. There are many other case attributes that we might investigate in determining generalizability, but these three pose obvious challenges to the Story Model. For example, criminal cases almost always involve sequences of goal-directed human activity (perfect for stories) but civil cases may also involve causal models other than stories, such as a mechanical causal model (e.g., Did the mechanic’s failure to tighten the lugnuts cause the accident?), an economic causal model (e.g., How much financial damage was caused to the company because of the trademark infringement?), or a biological causal model (e.g., Do the medical studies show that the industrial chemical was the cause of the employee’s disability?). There are also additional steps in civil case decisions which are not presently included in the Story Model, such as the need to establish a standard against which actual behavior can be compared to determine negligence, or the processing involved in translating suffering into damage amounts.

Within the criminal case domain, there are also case types that may not involve the central role of stories that we have proposed. For example, a case in which identification of the perpetrator (e.g., based on eyewitness testimony or extensive circumstantial evidence) is the central issue could involve less extensive story construction and more reasoning about non-story events such as police procedures or the abilities of eyewitnesses.

A final concern is that our case materials to date have always provided the material for at least two possible stories. In actual trials, however, the defense may defend by attacking the prosecution evidence and/or story rather than by presenting an alternative version of events. In this case, the juror has only one story to evaluate which may itself be very uncertain. This may alter the processing that occurs as compared to the situation where the juror has a story to reject as well. Investigation of the ways in which these and other attributes influence processing will allow us to begin to formulate principles that
define the boundaries of applicability of the Story Model across the domain of legal cases.

The second part of the generalizability question concerns the extent to which the cognitive processing strategies that we have outlined interact with various practices in actual trials. For example, we have shown that the order of evidence presentation can make large differences in the context of case materials that are somewhat difficult to understand. But, what effect does evidence presentation order (as related to stories) have on actual trial outcomes? What effects do strong stories presented by attorneys in the opening and/or closing statements have on the ease of story construction, the number of alternatives considered, et cetera? Our research program to date has focussed on the cognitive processing questions and not on the role that evidence presentation plays in actual trials.

B. Elaboration of Process Components

The distinction between mental processes and mental structures (mental representations) is an important one in information processing and cognitive science theories of mental activity. Mental representations refer to the form in which information is stored or organized. In our work, we have provided extensive evidence that comprehension, in the context of the juror’s decision making, results in a mental representation of the evidence that is a network of beliefs in memory, structured according to a “story schema.” However, knowledge about the nature of the mental structure provides only partial information about the nature of the mental processes that constructed that structure. For example, descriptions of the structure do not tell us when particular inferences are made, what kinds of knowledge structures are the “premises” of the inferences, or how and when contradictory inferences are resolved. In the legal decision making task, these questions are crucial for understanding how decision making occurs. By understanding the time at which inferences are made during the comprehension of evidence, the judge’s instructions, and subsequent decision making, we will know whether jurors construct single or multiple stories; what factors influence the point at which alternative stories cease to be considered; and the extent to which processing stages interact to produce a decision.

All of our discussions of the Story Model have assumed that a single “best” story emerges,79 and have treated the proposed process-

---

79 For some evidence on this issue, see D. Kuhn, N. Pennington & B. Leadbeater, Adult Thinking in Developmental Perspective, in 5 Life-Span Development and Behavior 157 (P. Baltes & O. Brim eds. 1983).
ing stages (story construction, verdict learning, and story classification) as occurring in sequence. Although the Story Model, as presently formulated, does not include stage interaction, it is possible that the story construction, verdict learning, and decision-making stages interact. For example, there is an indication in our previous data that one way jurors have of trying to understand verdict definitions is in terms of the “stories” they imply.\textsuperscript{80} This would be an obvious influence of story construction on verdict understanding. It is also possible that verdict understanding influences story construction by modifying a tentative story to fit more closely the constraints imposed by the verdict definitions. Moreover, we could find that a tentative decision influences further elaboration of the story, leading to a firmer decision. Although the Story Model as it now stands does not predict these interactions, other decision-making research has suggested that there exists a process of “bolstering” in which a tentative decision cycles back and promotes reconsideration and elaboration of the evidence.\textsuperscript{81} If this more cyclical view of decision making has merit (or if it is found to be false), it has implications for improving comprehension of the judge’s substantive instruction on the law, for the effects of variations in the timing of substantive instructions, and for the effects of including or excluding certain verdict categories as choice options.

C. Determinants of Confidence

A very important direction for development of the Story Model involves elaborating and formalizing the principles that we suggest determine confidence in decisions: coverage, coherence (completeness, consistency, and plausibility), uniqueness, and goodness-of-fit. One part of the puzzle involves examining these principles separately and in interaction. For example, in actual case materials, the coverage of the story and its completeness would often be correlated, although in principle they need not be. Experiments can examine the effects of these variables independently. Our empirical work so far has suggested that there are interactions we did not anticipate. For example, we suggested that uniqueness (the extent to which there is only one coherent story) will enhance confidence. Yet, we found that in one set of case materials, mock jurors were most confident when they were able to easily construct both stories. We suspect this was because one of the stories was less plausible (and thus less coherent) than the

\textsuperscript{80} See Pennington & Hastie, Evidence Evaluation, supra note 7.

other. On the other hand, it could be that knowing both stories will always increase confidence, regardless of plausibility. These are empirical questions that can be addressed through experimentation.

A second goal in pursuing determinants of confidence is to formalize these principles in order to understand how confidence in a decision can result from a computation across semantic features of a mental representation of evidence. Toward this end, we have adopted a formalization of the goodness-of-fit, coverage, coherence, and uniqueness ideas, based on Thagard’s model of explanatory coherence, ECHO. 82 We call our modification STORY-ECHO, for the obvious reason that we hypothesize that explanations of the evidence in legal trials take the form of stories. 83 The model is an interactive activation model that represents hypotheses and supporting evidence in a network of interconnected propositions. 84 Similar models have been implemented to describe comprehension processes and representations in non-decision-making tasks 85 and proposed for analogical and deductive reasoning tasks that are similar to our legal judgment task. 86 Although our modeling work is preliminary, we expect to be able to provide more explicit comparisons between our ideas about confidence and uncertainty and those proposed by probability calculi as applied to the legal decision task. 87

D. Stories as Mediators

Finally, there are many aspects of information in trials that we have not specifically incorporated into the Story Model. For example, we have not specified how some kinds of credibility information, such as source reliability and bias (consistency and plausibility aspects of credibility are part of the model) or other kinds of information commonly called “extra-legal,” will be processed or enter into story construction. 88

82 Thagard, Explanatory Coherence, 12 BEHAV. & BRAIN SCI. 435 (1982).
83 For a preliminary report on the work, see N. Pennington, P. Messamer & R. Nicolich, supra note 21.
85 E.g., W. KINTSCH, supra note 11.
87 For such a comparison, see P. Thagard, Probabilistic Networks and Explanatory Coherence 11 (rev. ed. Feb. 1990) (unpublished manuscript) (copy on file with author).
88 In citing extralegal information as a “kind” of information, we are not attempting to join either the descriptive or normative debates concerning rules of admissible evidence. The descriptive debate concerns whether or not such information actually affects decisions. There is ample evidence that it does and the task for our theory is to be able to predict the conditions under which it will or will not. The normative debate concerns whether such information
There is a large literature documenting the influence of information that is available at trial but has been called “extralegal” because it is either not allowed as a consideration in the decision or is allowed for limited purposes.\textsuperscript{89} For example, a number of studies have investigated the influence of the defendant’s prior criminal record, the moral character of the defendant, attitude similarity between the defendant and the juror or between the victim and the juror, the personal attractiveness of the defendant, and so forth. In general, the effects of these sources of information are complex, depend on the specific cases, and interact with one another. Dane and Wrightsman note: “To date, no single theoretical approach has been successfully applied to the full range of effects. . . . The lack of an integrative approach may well have contributed to the plethora of research findings.”\textsuperscript{90}

Up to the present, in developing the Story Model, we have not tried to account specifically for these effects of extralegal information. Our basic claim is that story construction mediates the effects of information of all types, when stories can be constructed. Therefore, our predictions are that extralegal information will have substantial effects on decisions under two conditions and not under a third. First, we predict that when it is difficult to construct a story to summarize the evidence, then the extralegal information will have a substantial effect (i.e., not be mediated by a story because a story cannot be constructed). This is equivalent to a situation where there is little “information” on which to base an opinion. Under these conditions, we predict that the effects of particular pieces of information will be direct; especially information, such as extralegal information that can be construed as “bad” or “good” even in the absence of an interpretive framework. Second, we predict that extralegal information will have moderate effects in the direction of the information when it is information that is related to and consistent with the story being constructed; when it is contradictory, we expect it to have the opposite effect. Finally, when the extralegal information is unrelated to the story being constructed, but the stories are easy to understand, then the extralegal information will have little impact on decisions.

These complex predictions are speculative but are consistent with


\textsuperscript{90} Dane & Wrightsman, \textit{supra} note 89, at 87.
a model developed by Petty and Cacioppo\(^1\) in which they suggest that persuasive information can be influential via two routes: one in which constructive processing takes place and one in which little "reasoning" at all occurs. In the latter, people are persuaded by superficial aspects of persuasive messages such as the social status of the speaker.

### E. Stories in Group Deliberation

We have proposed that individuals construct interpretations of evidence that combine the information heard and seen at trial with their knowledge of the world and their understanding of how a "good" story behaves. Jurors who construct different stories will either have brought different bases of world knowledge to the task or will have incompletely processed information presented at trial. It seems obvious that one effect of group deliberation is that different experiences with the world can be shared, thus opening up alternate interpretations for some jurors.\(^2\) An example of this from our videotapes of deliberating mock jurors is the interpretation of what it means to carry a knife. Some jurors (usually from "better" residential neighborhoods) would not at first accept that a person could carry a knife without intending to do specific harm with it. Other jurors pointed out during deliberation that either they or members of their families regularly carried knives because of the neighborhoods they lived or worked in, even though they had no specific intentions of using the knives.

Although sharing knowledge in deliberation is likely to open ways for people to change their stories, we have speculated on conditions that make it difficult for people to alter their accounts of "what happened." For example, there is limited evidence suggesting that once a juror has committed to a decision, it becomes more difficult for that juror to change stories.\(^3\) This suggests that deliberation strategies which combine maximal sharing of information and minimal early commitment to decisions will ensure the broadest coverage of evidence and relevant background knowledge.

### Conclusion

We have conducted a long series of investigations on the Story Model and believe that it is an excellent method for explaining and

---


\(^3\) *Id.* at 100.
predicting juror decision making in criminal trials. Clearly there are many areas for further theoretical development and empirical research. Perhaps the most satisfying characteristic of the Story Model approach for us, as cognitive experimental psychologists, is the extent to which it connects important naturally occurring decision-making phenomena to accounts from the mainstream of modern information processing theories of the mind.
Prototypes in the Courtroom: Lay Representations of Legal Concepts

Vicki L. Smith
Northwestern University

To select a verdict, jurors must integrate the evidence presented at trial with the law to select a legally appropriate verdict. Because most jurors have not had formal training in legal decision making (American Bar Association, 1968), the judge must instruct the jurors on the law at the time of trial. In a criminal trial, these instructions cover three types of law: (a) procedural law, including how to assess witness credibility and how to evaluate different kinds of evidence; (b) the requirements of proof, which explain that the defendant is innocent until the State proves beyond a reasonable doubt that the defendant is guilty; and (c) the substantive law, including definitions of the crimes charged (e.g., murder, robbery, and assault) and any defenses proposed (e.g., self-defense). For each crime, the law specifies a set of features that are singly necessary and jointly sufficient for guilt. Burglary, for example, requires that (a) the defendant knowingly entered a building, (b) the defendant did so without authority, and (c) the defendant did so with the intent to commit a crime (American Law Institute, 1962). Each of these criteria is a necessary condition for guilt; in other words, if any one of these conditions is not met beyond a reasonable doubt, then the juror is supposed to vote not guilty. These features are also jointly sufficient for guilt; if all three criteria are met beyond a reasonable doubt, then the juror is supposed to vote guilty.

The judge's instructions on the law are typically delivered at the end of the trial, after the evidence has been presented and immediately before deliberation. So, jurors first listen to days or weeks of testimony, then receive the judge's instructions on the law, and then must integrate facts and law to select a verdict. Efforts to understand how jurors accomplish this task have focused primarily on the evidence-evaluation stage of decision making, and several models of this process have been developed. Most of these models propose that jurors assess the meaning of each piece of evidence for the defendant's guilt and then mathematically combine the evidence items into a summary estimate of the probability that the defendant is guilty (Kaplan, 1977; Marshall & Wise, 1975; Pennington & Hastie, 1981; Thomas & Hogue, 1976). The information integration model, for example, proposes that jurors assign each piece of evidence a scale value corresponding to how incriminating the evidence is and a weight representing how important it is for the determination of guilt. Jurors then compute a weighted average of the scaled evidence items to generate an estimate of the probability that the defendant is guilty (Kaplan, 1977; Ostrom, Werner, & Saks, 1978; Pennington & Hastie, 1981). The Bayesian model proposes that jurors combine their prior beliefs about the defendant's guilt with subjective likelihood ratios that estimate the probability of guilt conveyed by each piece of evidence. These elements are then multiplied to generate a posterior probability estimate of the defendant's guilt (Marshall & Wise, 1975; Pennington & Hastie, 1981; Schum & Martin, 1982).

These and other mathematical models of juror decision making differ in the particular assessments jurors are proposed to make about the evidence and the mathematical rules they use to combine evidence items, but they are similar in their focus on the evidence-evaluation stage of decision making. The
only elements of the law incorporated in these models are the presumption of innocence, which is modeled as jurors’ initial beliefs about the defendant’s guilt, and the reasonable doubt standard, which is the decision threshold against which the juror’s final probability estimate is compared to select a verdict. If the juror’s estimate of guilt exceeds this threshold, then the juror votes guilty; if not, then the juror votes not guilty. The mathematical models are vague or altogether silent about how jurors make sense out of the procedural and substantive law the judge presents and about how that law is integrated with the facts to select a verdict.

A more recent model of juror decision making proposes that jurors make sense of the evidence presented at trial by constructing a story of what happened (Bennett, 1978, 1979; Pennington & Hastie, 1986, 1988). This story is based on the testimony of the witnesses and on the jurors’ inferences about the evidence from their general knowledge of the physical and social world. Jurors then listen to the judge’s instructions on the law and generate a list of critical features representing the verdict categories. They compare these features to the attributes of their stories and select the verdict that best matches the story that they have constructed (Pennington & Hastie, 1986, 1988). Although the story model is more complete in its analysis of the juror’s task, including the representation of the law and the integration of facts and law in addition to the evaluation of evidence, research on the story model has focused primarily on story construction (Pennington & Hastie, 1986, 1988). As yet, very little is known about how jurors represent verdict categories, how they use the information contained in the judge’s instructions, or how they integrate facts and law to select a verdict. The purpose of the present research was to investigate these gaps in our understanding of how the law influences jurors’ decision making.

Jury Instructions

The judge’s instructions are intended to educate formally untrained jurors in the legal concepts that apply to the case that they must decide. However, research on jury instructions has revealed that jurors’ understanding of the law is quite poor. In some experiments, mock jurors performed at chance levels on tests of comprehension for the instructions (Elwork, Sales, & Alfani, 1977; V. L. Smith, 1991). In others, instructed jurors performed no better than uninstructed jurors (Elwork et al., 1977; Severance & Loftus, 1982; V. L. Smith, 1991). Some researchers have argued that it is the complex language of the instructions that impedes understanding, and a few experiments have demonstrated that rewriting the instructions in simpler language can significantly improve jurors’ comprehension (Charrow & Charrow, 1979; Elwork et al., 1977; Severance & Loftus, 1982). Other researchers have targeted the procedural aspects of instruction. Jury instructions are delivered orally by the judge after the evidence has been presented. Reformers propose that instructing the jury before the trial, allowing jurors to take notes on the instructions, or giving them written copies of the instructions to take into deliberation might assist jurors in their efforts to learn the law (Goldberg, 1981; Prettyman, 1960; Sand & Reiss, 1985). Some research indicates that presenting the instructions both before and after the evidence (rather than only after) improves jurors’ comprehension (V. L. Smith, 1991) but that allowing jurors to take notes or giving them written copies of the instructions have not shown such benefits (Hastie, 1982; Heuer & Penrod, 1988, 1989).

Although rewriting and pretrial instruction have improved jurors’ understanding of the law, they have not solved the problem of inadequate comprehension. There are still important deficits in understanding even when these procedures are implemented. What else might contribute to jurors’ difficulty? Jury researchers and legal commentators (McBride, 1969; Prettyman, 1960) have more or less implicitly treated jurors as blank slates on which the law is written; it is assumed that jurors’ only source of information about the law is the judge’s instructions and that the aim of instruction is to create legal concepts where none exist. Perhaps, however, jurors bring to trial some knowledge of the law gained from informal sources such as newspapers, television, and movies. These naive concepts may serve as another source of information about the law and may influence jurors’ processing and use of the judge’s instructions.

There are three possible states of jurors’ prior knowledge of the law. First, jurors may bring no information about the law to trial; learning the necessary concepts from the judge’s instructions. This is consistent with the characterization of jurors as blank slates. The goal of jury instruction, then, would be to create new and accurate legal concepts. Second, jurors may bring some correct information about the law to trial. In this case, the instructions must fill the gaps in jurors’ existing concepts. Mock jurors’ chance performance on tests of comprehension suggests that people actually have little correct information about the law. The fact that instructed jurors generally perform no better than uninstructed jurors also suggests that the instructions do not succeed in filling the gaps in people’s knowledge. The third possibility is that jurors bring some incorrect information about the law to trial. In this case, jurors must learn from the judge’s instructions that their existing concepts are wrong, and they must learn the correct concepts. The goal of jury instruction, then, would be concept revision rather than concept formation; jurors’ misconceptions must be replaced by an accurate representation of the law. This goal may require a different approach to jury instruction; the process of revising an existing concept may be quite different from the process of creating a new concept.

What, then, do laypeople know about the law, and how is that information used? These questions have not been addressed in previous research on jury instructions, but if one considers that choosing a verdict is a categorization task, there is substantial research in both cognitive and social psychology that may shed light on this issue. When a juror is asked to decide whether a defendant is guilty or not guilty of a given crime, the juror must make a category membership decision. Do the defendant’s actions qualify as a member of the crime category charged against the defendant? A guilty verdict in an assault case, for example, means that the juror believes the defendant’s behavior is an instance of the category assault. A not-guilty verdict means the juror rejects the category label assault. So, a guilty/not-guilty verdict decision is essentially a yes/no category membership decision. How, then, do jurors make these categorizations?
Categorization

Although little is known about people's categorization of crimes, there has been considerable research on the categorization of concrete objects (Rips, Shoben, & Smith, 1973; Rosch, 1975, 1978; Rosch & Mervis, 1975; E. E. Smith & Medin, 1981). People's representations of common objects (e.g., animals, vehicles, or furniture) were traditionally thought to include a set of defining features that are singly necessary and jointly sufficient for category membership. If an object possessed all of the necessary features of the category, it would be judged a category member. If any necessary feature was not present, the object would be rejected as a category member (see E. E. Smith & Medin, 1981). However, substantial empirical evidence now indicates that people do not base their categorization decisions on a set of defining features, as suggested by the classical model (Rosch, 1973, 1975, 1978). Rather, people's knowledge of these object categories appears to be represented by a prototype, or typical exemplar, of the category. If a target object has many features that are characteristic of the prototype, then it will likely be judged a category member; if the target has few characteristic features, then it will likely be rejected as a category member. These features are associated with the category to varying degrees but are not necessary for membership. One implication of this prototype model is that categories have a graded structure; the more typical features an object possesses, the more representative of the category it is perceived to be. Consistent with this reasoning, the typicality of the target object has been found to influence both the speed and accuracy of categorization decisions (Rips et al., 1973; Rosch, 1973, 1975, 1978).

More recent research has demonstrated that this prototype model of categorization generalizes to abstract, social categories as well. Categories as diverse as social situations (e.g., parties, interviews, and work), personality traits (e.g., shy and dominating), and social groups (e.g., homosexuals and fraternity members) all appear to be organized around category prototypes (Buss & Craik, 1983; Cantor, Mischel, & Schwartz, 1982; Lord, Lepper, & Mackie, 1984; Rothbart & Lewis, 1988). These categories exhibited a graded structure that was quite reliable across subjects (Buss & Craik, 1983; Cantor et al., 1982) and that influenced subjects' behavior toward group members (Lord et al., 1984) and their inferences about the group (Rothbart & Lewis, 1988). This research indicates that prototype representations operate in a variety of decision contexts, both concrete and abstract, both physical and social.

Given people's facility with constructing and using prototypes, it is possible that they have generated similar representations of legal concepts as a way of organizing the informal and fragmentary information they encounter about the law. They may, for example, believe that the typical burglary is committed by a person dressed in black clothing who breaks into an unoccupied house late at night and steals jewelry. If laypeople have constructed prototypes of crime categories, then their naïve decision strategy may reflect the graded structure of the category. They may determine the guilt or innocence of a defendant by comparing the characteristics of the defendant's crime with the features of their prototype. Guilty verdicts would result if there was sufficient feature overlap; not-guilty verdicts would result if the correspondence was low. This strategy is clearly at odds with proper legal decision making, in which guilt is based on the satisfaction of a set of necessary and sufficient conditions.

One could argue that laypeople are unlikely to construct prototypes of crime categories because they are unnecessary. The judge specifies the defining criteria for the crime in the instructions to the jury, so the decision task is relatively well-defined. Perhaps a preference for prototype-based decision making emerges only in social contexts that lack a well-specified decision process. Previous social categorization experiments have focused on categories or decision tasks that are relatively poorly defined. Personality traits (Buss & Craik, 1983) and social situations (Cantor, Mischel, & Schwartz, 1982) do not have clear category boundaries, but crime categories do. There is no normatively correct strategy for deciding whether one would like to work with a target person (Lord, Lepper, & Mackie, 1984) or for inferring how politically liberal members of a target group are (Rothbart & Lewis, 1988), but there is a correct strategy for verdict selection, and that strategy is specified in the judge's instructions.

The present research considers whether people have constructed and use prototype representations in the legal context when a competing strategy is required. This research begins to fill the gap in our understanding of how jurors use legal concepts for verdict selection. More generally, this research considers how flexible people are in their choice of decision strategies. If jurors have constructed prototypes of crime categories, are they flexible enough to set aside their preferred strategy and use an alternative decision process when instructed to do so? To answer these questions, the following studies investigate (a) whether people have naive concepts of crime categories and, if so, whether those concepts contain prototypes; (b) whether people's naive representations influence their verdict choices; and (c) how the judge's instructions influence people's decision strategy.

Experiment 1

Experiment 1 examines whether people have naive concepts of crime categories and, if so, whether those concepts contain defining features (consistent with the law) or associated features (consistent with a prototype representation). In Part 1, subjects were asked to list features that they considered to be characteristic of several different crime categories (Buss & Craik, 1983; Cantor, Mischel, & Schwartz, 1982; Hampton, 1981; Rosch & Mervis, 1975). Part 2 investigates whether subjects used these characteristic features as necessary conditions for guilt or merely as features associated with guilt. Part 3 considers whether the characteristic features constitute a category prototype. If so, then subjects' concepts should reveal a graded structure, with some category members perceived as more typical than others.

Part 1

To determine whether people have compiled information about various crimes that they can access for decision making, subjects were asked to list all of the features that they believed characterize different crime categories. If subjects have con-
constructed a representation, then they should be readily able to list its features.

Method

Subjects and procedure. Fifty-two introductory psychology students at Northwestern University participated in partial fulfillment of a course requirement. The experimenter told subjects that they would be asked to list all of the features and attributes that they felt were common to, or characteristic of, different kinds of crimes. For each crime, they would have 2 min to write down all of the attributes that they could think of. They were instructed to not simply free-associate to the crime name but to list features that were common to the category in general.

Materials. Each page of the questionnaire had a different crime name printed at the top. Vandalism and rape, the first two crimes listed, were intended as practice. Crimes on the remaining 10 pages were randomly ordered for each subject to minimize the effects of additional practice or fatigue. After completing their lists for the 12 crimes, subjects were fully debriefed. Feature lists were solicited for 10 crimes, but only 5 were targeted for analysis at this time: assault, burglary, kidnapping, murder, and robbery. These crimes were selected for two reasons. First, they represent a wide variety of crime types, including both crimes against the person and crimes against property and both violent and nonviolent crimes. Second, they are among the 10 most frequently listed members of the superordinate category crime (Battig & Montague, 1969). As a result, they are likely to be familiar to subjects and to provide a good test of alternative representational structures.

Results and Discussion

Subjects had no trouble listing the features they associated with these crimes, and they did not ask for clarification of the instructions. The average subject provided about seven features for each crime, including the characteristics of the victim (e.g., child, injured, or scared) and the perpetrator (e.g., armed, disguised, or poor), the location of the incident (e.g., home, playground, or store), the event (e.g., threat, physical attack, or stabbing), and the reasons for committing the crime (e.g., revenge, anger, or money). Apparently, subjects had constructed naive representations of these crimes that they could access when asked to do so.

A master list of features was compiled for each crime category, and the number of subjects listing each feature was computed. When computing these frequencies, features were considered equivalent if they matched exactly or if they were synonymous. For example, many subjects indicated that burglaries are committed when the victim is absent from the scene of the crime, as in each of the following: “done . . . while victim is away from property,” “victims are not home,” and “done when victim is away . . . victim is not present.” Features that were not synonymous were considered different features. For example, sexual attack and physical attack are related to each other but were considered different features of assault because they are not synonymous.

Across the five crimes, very few features were listed consistently by subjects. No feature was listed by all subjects, and only one or two features for each crime were provided by as many as half of the subjects. Table 1 shows the features that were listed by 30% or more of the subjects for each category. As indicated in the table, 90% of the subjects said that assault involves a physical attack, 63% said that kidnapping involves a ransom demand, and so on. The third column of Table 1 lists the defining criteria for each crime as specified by the law. The complete definition of each crime, as it appears in the Illinois Pattern Jury Instructions (1981), is reproduced in Appendix A. The defining features specified by the Illinois instructions also appear in the Model Penal Code definitions of these crimes, which have been used as a guide by many states when drafting their criminal codes (American Law Institute, 1962; see also Black, 1983). As illustrated in Table 1, a considerable discrepancy exists between the features subjects listed and the correct defining features of these crimes. Robbery, for example, involves the taking of property from the victim by force or threat of force, but robbery does not require that the property be valuable, that the perpetrator be armed, or that the location be someone’s home. In general, then, subjects did not provide the correct defining features on their feature lists. Nevertheless, it is possible that these frequently listed features operate as necessary conditions in subjects’ representations, even though the features are legally incorrect. Part 2 was designed to investigate this possibility.

Part 2

If the frequently listed features from Part 1 constitute defining features, then they should operate as necessary conditions for guilt. In contrast, if these features constitute a category prototype, then they should be associated with the category but would not be necessary conditions for guilt. To determine whether a given feature X was a necessary condition for guilt, a target crime was constructed that did not contain feature X but did contain many other characteristic features of the category. If X is a necessary condition, then subjects should reject the target as a category member and vote not guilty because X was not present. However, if X is merely associated with the category and is not necessary, then subjects should judge the target to be a category member and vote guilty because the target contained many characteristic features of the category, even though X was not present. On the basis of this reasoning, several crime scenarios were constructed to test whether the features appearing in Table 1 were necessary conditions for guilt or were merely associated with category membership.

Method

Subjects and procedure. Forty-one introductory psychology students at Northwestern University participated in partial fulfillment of a course requirement. The experimenter told subjects that they would select verdicts for several crime situations and that they should base their verdicts on what they knew about each crime. Subjects completed the questionnaire (described below) at their own pace and were debriefed.

Materials. For each crime category, subjects read short scenarios (five sentences on the average) describing a criminal encounter. These scenarios included as many features as possible from subjects’ feature lists, so that the scenarios would be perceived as characteristic of the category. The scenarios systematically excluded the most frequently listed features appearing in Table 1 to determine whether any of these features operated as a necessary condition for guilt. For example, one of the robbery scenarios read as follows:
This scenario omitted two frequently listed features from the robbery category: (a) money is taken and (b) occurs in a home or apartment. However, the facts included two other frequently listed features: (a) the incident was planned, (b) it occurred at night, (c) it occurred in a public place, and (d) the perpetrator was a man. This fact situation, then, included many characteristic features of the robbery category but omitted two frequently listed features to determine whether those features were necessary for guilt. If either of the two omitted features is necessary, then subjects should reject the defendant’s actions as a member of the robbery category and vote not guilty. However, if these two features are merely associated with robbery but are not necessary for guilt, then subjects should vote guilty because the scenario contains many other characteristic features of the crime.

Fourteen critical scenarios tested whether the features listed in Table 1 were necessary conditions for guilt: 1 assault, 3 burglaries, 4 kidnappings, 2 murders, and 4 robberies. Subjects’ questionnaires contained 7 critical and 5 filler scenarios (1 of each crime type). For the critical scenarios, half of the subjects made verdict decisions for 1 assault, 1 burglary, 2 kidnappings, 1 murder, and 2 robberies; the other half chose verdicts for 2 burglaries, 2 kidnappings, 1 murder, and 2 robberies. In this way, each critical scenario was presented to at least 20 subjects. All of the subjects read the same 5 filler crimes. The fillers were designed to be uncharacteristic of the category in question and were expected to produce not-guilty votes. There were two versions of each questionnaire, each with a different random combination of critical scenarios. There were no significant differences in response patterns for the alternative versions of the questionnaire, so the data were collapsed across versions.

Results and Discussion

The proportion of guilty votes was computed for each scenario and tested against zero. If a target feature is necessary, then subjects should vote not guilty when the feature is not present. To the extent that the conviction rate exceeds zero, subjects are not using the omitted feature as a necessary condition. Across the 14 critical scenarios, the conviction rates ranged from 71% to 100%, with a mode of 100%. All of the conviction rates significantly exceeded zero (all ps < .0001), indicating that none of the frequently listed features was a necessary condition for guilt.1 It appears, then, that these attributes

---

1 The conviction rate data were analyzed using a z test for binomial proportions. This test cannot tolerate a zero performance criterion because the error term would then be zero and z would be undefined. To avoid this problem, a value close to zero (01) was selected as the criterion. It can be argued that zero is an unrealistically stringent performance standard. There are few nontrivial tasks on which subjects respond uniformly. Because this was a two-alternative forced-choice decision, the conviction rate would have been 50% if subjects had disregarded the facts and chosen a verdict at random. It is, then, more conservative to use a 50% performance criterion. The conviction rates for all of the scenarios significantly exceeded 50%, supporting the conclusion that none of the frequently listed features was necessary for guilt.
are associated with the crime categories but do not operate as defining features. Part 3 was designed to test the possibility that these features constitute category prototypes.

**Part 3**

If subjects' representations contain prototypes, then their categories should demonstrate a graded structure; fact situations possessing many characteristic features of the category should be perceived as more typical than fact situations containing few characteristic features. This kind of graded structure does not emerge in classically defined categories, in which a target is not a category member if it does not possess the appropriate necessary conditions. For classically defined categories, all category members are equally representative, because all meet the same membership criteria (E. E. Smith & Medin, 1981). To determine whether people's representations of these crimes exhibit a graded structure, subjects judged the typicality of several short fact situations. All of these scenarios met the legal requirements for the crime in question and were thus category members, but they differed from each other in the number of characteristic features (from Part 1) that they included.

**Method**

**Subjects and procedure.** One hundred Northwestern University undergraduates were paid $5 each for participating in this experiment. Subjects were told that they would be asked to judge how typical various crime situations are of certain crime categories. To illustrate what was meant by typicality, the experimenter explained that certain shades of red strike us as more typical of the color red than others, and some breeds of dogs are better examples of the dog category than others. (These instructions were adapted from Rosch & Mervis, 1975.) Similarly, in this experiment subjects would rate how typical they thought each fact situation was of the crime category specified.

**Materials.** There were six scenarios for each category, all of which met the legal requirements for the crime in question (see Appendix A). The kidnapping scenarios, for example, all involved defendants secretly confining the victims against their will. The scenarios differed in how many characteristic features of kidnapping they included. (See Appendix B for examples; for each scenario, the underlined segments correspond to characteristic features from the Part 1 feature lists.) For each category, two scenarios contained many characteristic features (M = 13 per scenario), two contained an intermediate number of these features (M = 10), and two contained relatively few characteristic features (M = 8). The lengths of the scenarios within each category were equivalent, so differences in perceived typicality could not be due to differences in scenario length. The average length was 6 sentences, or 94 words. Subjects rated the typicality of the scenarios for each category separately. The questionnaires listed all six scenarios for the crime on two or three pages, with the name of the category typed prominently at the top of each page. When subjects finished rating the scenarios of the first category, they set those aside and rated the scenarios of the second category, and so on through all five categories. Ratings were made on 7-point scales, ranging from not at all typical (1) to very typical (7). There were three random orders of scenarios within each category. No order effects emerged, so the data were collapsed across orders. The order of categories was randomly determined for each subject. After completing all 30 scenarios, subjects were debriefed.

**Results and Discussion**

If subjects' representations of these crime categories contain prototypes, then scenarios incorporating many characteristic features should be rated significantly more typical than scenarios incorporating few such features. In contrast, if the defining features are what matter for judgments of typicality, then subjects' ratings should not differ across scenarios, because all of the scenarios are category members. For each crime category, the mean typicality rating was computed for scenarios with many, intermediate, and few characteristic features (see Table 2). Repeated measures analyses of variance (ANOVAs) revealed significant differences across the three scenario types for each crime (all ps < .0001). Planned, matched-pairs t tests indicated that the scenarios containing many characteristic features were rated significantly more typical than scenarios containing few features for all five crimes (all ps < .0001). Intermediate-level scenarios were rated significantly more typical than few-featured scenarios for all five crimes (all ps < .0001). Many-featured scenarios were perceived as more typical than intermediate-level scenarios for burglary, kidnapping, and murder (all ps < .0001) but not for robbery and assault.

Also of interest was the reliability of subjects' typicality judgments. Correlations of typicality ratings were computed across all possible pairs of subjects for each crime. So, for assault, the typicality ratings of Subject 1 were correlated with those of Subjects 2, 3, . . . , 100; the typicality ratings of Subject 2 were correlated with Subjects 3, 4, . . . , 100, and so on. These correlations were averaged, and the average was used to compute Cronbach's alpha as an index of reliability. This analysis was performed for each crime. The average correlations ranged from .25 for assault to .60 for burglary (see Table 2). The alpha coefficients revealed that the typicality ratings were extremely reliable across subjects (all as > .96).

These findings are consistent with category representations based on prototypes. Fact situations that contained many characteristic features were perceived as significantly more typical than those containing few characteristic features, indicating that subjects' representations have a graded structure. Furthermore, there was considerable agreement on this internal structure across subjects; perceptions of what is typical and what is atypical were extremely reliable.

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Mean Typicality Ratings and Reliability Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of characteristic features</td>
</tr>
<tr>
<td>Crime category</td>
<td>Many</td>
</tr>
<tr>
<td>Assault</td>
<td>4.53a</td>
</tr>
<tr>
<td>Burglary</td>
<td>6.02a</td>
</tr>
<tr>
<td>Kidnapping</td>
<td>5.75a</td>
</tr>
<tr>
<td>Murder</td>
<td>5.58a</td>
</tr>
<tr>
<td>Robbery</td>
<td>5.25a</td>
</tr>
</tbody>
</table>

Note: Numbers with different subscripts in a row differ significantly (p < .0001). Numbers with identical subscripts in a row do not differ significantly (p > .10).
These data do not rule out the possibility that subjects’ representations of these categories contain one or more necessary features that have not yet been identified. In Part 1, subjects were asked to describe their representations on a feature-listing task, and it is quite possible that this task did not produce an exhaustive list of features from all subjects. There may have been some attributes that subjects felt were too obvious to write down or that they failed to list for some other reason. Most people, for example, probably believe that murder involves a dead body, but this feature was listed by only three subjects. Such features probably exist in other categorization domains as well. People probably require a target object to be inanimate before they will label it a chair, but they may not volunteer this feature when asked to describe their representation of the chair category. However, the possible existence of one or more as yet unidentified necessary conditions does not mean that people have not constructed prototypes of these crime categories. The goal of Experiment I was not to generate an exhaustive description of people’s representations; it was to identify the features people readily associate with the categories and to determine how those features influence people’s judgment and decision making. Subjects in Experiment I volunteered many characteristic features of these crimes, the most frequent of which were not necessary for guilt. The presence or absence of these same features was varied in Part 3, and these variations influenced subjects’ ratings of typicality. Apparently, then, these features are important for subjects’ perceptions of fact situations, and this poses problems in a legal setting in which such features are supposed to be irrelevant as decision criteria. The possibility that one or more necessary conditions exist that have not yet been discovered does not change this conclusion.

In summary, it appears that people construct prototype representations of crime categories, similar to the way other social categories are organized (Buss & Craik, 1983; Cantor, Mischel, & Schwartz, 1982; Lord, Lepper, & Mackie, 1984; Rothbart & Lewis, 1988), but contrary to the way the categories are organized under the law. The next step is to determine the importance of these prototypes for decision making. Does the perceived typicality of a fact situation influence people’s verdict choices, or is it disregarded for categorization decisions? Perhaps people have constructed prototypes for these crimes, and we can demonstrate their existence, but they do not affect people’s verdict choices.

**Experiment 2**

Experiment 2 tested the influence of subjects’ prototypes on their category membership decisions in two ways. In Part 1, subjects read several short fact situations and were asked to determine what crime the defendant committed. Thus, subjects were asked to generate the appropriate category label on the basis of their own knowledge of these crimes. In Part 2, subjects were asked to determine whether the defendants were guilty or not guilty of a particular crime. So for this task, the category label was provided, and subjects had to determine whether to accept the label (and vote guilty) or reject the label (and vote not guilty). If these categorization decisions are influenced by subjects’ prototypes, then scenarios containing many characteristic features of the category (typical crimes) should be judged category members more often than scenarios containing few characteristic features (atypical crimes).

**Part 1**

In this experiment, subjects read the typical and atypical fact situations from Experiment I (Part 3) and were asked to determine what crime the defendant committed. If the typicality of the crime influences decision making, then subjects should provide the correct category name more often for typical than atypical instances of the category. If the defining features are what matter for categorization, then subjects should correctly label all category members.

**Method**

**Subjects and procedure.** Forty-six introductory psychology students at Northwestern University participated in partial fulfillment of a course requirement. The experimenter told subjects that they would read several short fact situations and decide for each what crime the defendant had committed, on the basis of what they knew about the law. If they were unsure of the correct answer, then they were to make their best guess. Subjects completed the questionnaires (described below) at their own pace and were debriefed.

**Materials.** The two typical and two atypical crime scenarios for each crime category from Experiment I (Part 3) were used as stimuli (20 scenarios total). Appendix B contains examples of the typical and atypical assault, kidnapping, and robbery scenarios. The atypical scenario used in these experiments had to be constructed particularly carefully. To be guilty of assault, the defendant must place the victim in reasonable fear of bodily harm. The feature lists from Experiment I (Part 1) suggested that most people confuse assault with battery, which involves actual injury to the victim. The most frequently listed features of assault were “physical attack” and “victim is injured.” To ensure that this category confusion did not account for subjects’ typicality ratings or their verdict decisions, all of the atypical scenarios used in these experiments met the legal requirements for assault but not the legal requirements for battery. The defendants placed the victims in reasonable fear of bodily harm but did not physically attack or injure their victims.

Subjects read each scenario and were asked (1) what is the most serious crime (defendant) is guilty of?; and (2) Is (defendant) guilty of any other crime(s)? If so, what are they? The name of the particular defendant was inserted for each scenario. There were three versions of the questionnaire, each with a different random order of scenarios. No order effects emerged, so the data were collapsed across versions.

**Results and Discussion**

If subjects’ prototypes affect their categorization decisions, then they should volunteer the correct category name more often for the typical than for the atypical scenarios. The number of subjects who provided the correct label as the defendant’s most serious crime was computed. Any response that included the category name was considered correct. For example, “premeditated robbery” was counted as a correct response for the robbery scenarios, “attempted armed assault with intent to harm” was considered correct for the assault scenarios, and so on. The correct category name was provided significantly more often for typical than atypical scenarios for four of the five
crimes: assault, 55% versus 34%, \(x^2(1, \ N = 46) = 8.45, p < .01\); burglary, 22% versus 0%, \(x^2(1, \ N = 46) = 22.22, p < .001\); kidnapping, 91% versus 24%, \(x^2(1, \ N = 46) = 84.74, p < .001\); and robbery, 76% versus 29%, \(x^2(1, \ N = 46) = 40.32, p < .001\). For murder, almost all of the subjects said the defendant was guilty of murder, regardless of the typicality of the scenario, 88% (typical) versus 87% (atypical), \(x^2(1, \ N = 46) = 0.04, ns^2\).

There are two possible reasons why typicality did not influence the categorization of the murder scenarios. One possibility is that subjects knew and used the correct defining features of murder. All of the scenarios were category members, and subjects may have made the right decisions for the right reasons. However, no evidence in the data from Experiment 1 indicates that subjects do know the legal definition of murder. The features that appeared most frequently on subjects' feature lists did not contain any necessary conditions, either correct or incorrect. Furthermore, if subjects knew and used the correct defining features of the murder category, their typicality ratings should not have varied across scenarios; all of the scenarios were category members and should have been perceived as equally representative of the category if the defining features are what influence subjects' perceptions. This evidence is not dispositive, but it is suggestive that subjects are unaware of the correct defining features of the murder category. Another possible reason that subjects volunteered the word murder at such a high rate is that they had no other label readily available. Manslaughter and homicide are alternative names for killings, but these are less common terms than murder; neither appears in the Battig and Montague (1969) norms for the category crime. In this experiment, only 4% of the subjects used the label manslaughter, and only 5% used the label homicide.

These data indicate that typicality does influence categorization decisions for assault, burglary, kidnapping, and robbery. For these crimes, subjects were better able to identify category members when the facts were typical of the category than when they were atypical. Part 2 was designed to assess the impact of typicality on subjects' forced-choice categorization decisions.

**Part 2**

Rather than providing a category label for a fact situation, a task that may be particularly difficult for abstract categories, subjects in this experiment had to decide whether a given fact situation was a member of a given crime category. This task is more similar to the one facing actual jurors. In a trial, jurors do not have to generate crime names; they must determine whether the defendant is guilty of the particular crime (or crimes) charged. If people's prototypes affect their verdict decisions, then defendants who commit typical crimes should be judged guilty more often than those who commit atypical crimes.

**Method**

**Subjects and procedure.** Sixty-eight introductory psychology students from Northwestern University participated in partial fulfillment of a course requirement. The experimenter explained that the subjects would read several short fact situations and make verdict decisions for each on the basis of what they knew about the crime charged. If they were not sure whether to vote guilty or not guilty, they were to make their best guess. Subjects completed the questionnaires (described below) at their own pace and were debriefed.

**Materials.** The two typical and two atypical scenarios for each crime were used as stimuli (20 total). All of the critical scenarios were category members, so the correct verdict for each was guilty. Four filler scenarios for each category were added to the questionnaire and were designed to provide a balance of guilty and not-guilty votes. In total, then, subjects made judgments on 40 scenarios: 2 typical, 2 atypical, and 4 filler crimes for each of five categories. Subjects indicated their verdict decisions by checking yes or no to the question: "Is (defendant) guilty of (crime)?" The name of the particular defendant and the particular crime were inserted for each scenario.

Subjects made verdict decisions for each crime category separately. The questionnaires listed all eight scenarios for a category on three pages, with the category name typed prominently at the top of each page. When the subjects completed their verdict decisions for the first category, they set those aside and judged the defendants in the second category, and so on through all five crimes. There were three random orders of scenarios within each crime category; no order effects emerged, so the data were collapsed across versions. The order of categories was randomly determined for each subject.

**Results and Discussion**

If subjects' prototypes influence their verdict decisions, then typical crimes should produce a higher conviction rate than atypical crimes. The proportion of subjects voting guilty for the typical and atypical scenarios was computed over all five crimes (see Table 3). A matched-pairs \(t\) test indicated that the conviction rate was higher for the typical than for the atypical scenarios, \(t(67) = 17.8, p < .0001\). When the five crimes were considered separately, typical scenarios produced a higher conviction rate for assault, burglary, and kidnapping (\(ps < .0001\)). Murder and robbery showed no effect of typicality; typical and atypical scenarios produced identically high conviction rates.

2 Additional analyses were conducted to include subjects' responses to the second question: "Is (defendant) guilty of any other crime(s)? If so, what are they?" For these analyses, subjects' responses were considered correct if the correct category name was provided on either the most serious crime or the other crimes questions. With this less stringent accuracy criterion, subjects did not need to know the relative seriousness of the crimes they listed. The results of these analyses were equivalent to the first: Typical scenarios produced significantly more correct responses than atypical scenarios for assault, burglary, kidnapping, and robbery, but typicality did not affect the number of correct responses for murder.
There are two possible reasons that typicality did not affect conviction rates for murder and robbery. First, perhaps subjects knew and used the correct definitions of these two crimes. The correct verdict for these scenarios was guilty, and subjects may have made the right decisions for the right reasons. Again, however, there is no evidence in the data from Experiment 1 that subjects do know the correct definitions of these crimes; subjects’ feature lists did not contain necessary conditions, and their judgments of typicality were influenced by the number of characteristic features a category member possessed, a finding that is consistent with the graded internal structure of prototype, not classical, representations. A second possible explanation for these high conviction rates is that there was a sufficient condition for categorization that appeared in both the typical and atypical scenarios. Perhaps the features on subjects’ feature lists were not necessary for categorization (as demonstrated in Part 2 of Experiment 1), but there may have been one or more sufficient conditions for guilt included in these lists. A review of the critical scenarios for murder and robbery revealed two features for each crime that appeared in both the typical and atypical scenarios and that were listed relatively frequently on subjects’ feature lists. For murder, these features were “violent killing” and “intentional killing.” For robbery, the features were “something of value is taken” and “victim is present.” If these features are sufficient to produce guilty verdicts, their presence in both the typical and atypical scenarios would account for the high conviction rates obtained in this experiment. The sufficiency of these features to produce guilty verdicts was tested in Experiment 3.

The results of Part 2 indicate that for assault, burglary, and kidnapping, the typicality of the fact situation influenced subjects’ verdict choices. Apparently, these naive decision makers were guided by their prototypes when selecting a verdict. Using prototype information is inconsistent with proper legal decision making, in which verdict decisions are based on a set of legally appropriate necessary and sufficient conditions. Perhaps, however, subjects relied on their prototypes only because they were not informed of the correct decision process, and they would disregard their prior knowledge once they heard the judge’s instructions on the law. This raises the issue of decision flexibility. Do we use typicality-based decision strategies only when there is no alternative available, or does this preference persist in the face of instructions to use a competing strategy? Experiment 3 considered the effectiveness of the judge’s instructions in overriding subjects’ use of prototypes for decision making.

Experiment 3

If the judge’s instructions effectively convey the appropriate decision strategy, then the verdict decisions of instructed subjects should no longer be influenced by the typicality of the fact situation. All category members should produce guilty votes, regardless of how typical or atypical the facts are perceived to be. In this experiment, half of the subjects made verdict decisions without hearing the legal definition of the crime charged. They heard only the preliminary instructions that are delivered in all criminal trials, which explain that their verdicts should not be based on sympathy or prejudice, that they should consider the evidence in light of their own observations and experience in life, and that each defendant is presumed innocent until the juror is convinced beyond a reasonable doubt of the defendant’s guilt (Illinois Pattern Jury Instructions, 1981). Because these preliminary instructions do not define the crime charged, they were not expected to alter subjects’ decision strategies. They might make subjects more cautious about convicting the defendants and thus produce consistently lower conviction rates than were obtained in Experiment 2, but subjects should still convict significantly more often for typical than atypical crimes. The other half of the subjects in Experiment 3 heard both the preliminary instructions and the substantive instructions defining the crime charged (Illinois Pattern Jury Instructions, 1981; see Appendix A). These instructions specify the necessary and sufficient conditions for the crime and explain how to use these criteria; if all of the defining criteria are met beyond a reasonable doubt, then the juror should vote guilty, but if any criterion is not met beyond a reasonable doubt, then the juror should vote not guilty. If these instructions are effective, then subjects should convict equally often for typical and atypical scenarios, and the conviction rates should be near 100%, because all of the critical scenarios are category members.

Of secondary interest in this experiment was clarifying jurors’ decision strategies for the murder and robbery categories. Specifically, are there features in these scenarios that operate as sufficient conditions for guilt? Several new scenarios were written to test the potentially sufficient features identified in Experiment 2. These scenarios (a) contained few of the characteristic features from the feature lists of Experiment 1 so that they would be perceived as relatively atypical of the category and (b) contained one of the potentially sufficient conditions. If the target feature is sufficient for guilt, then subjects should vote guilty even though the scenario is otherwise atypical. If the target feature is not sufficient, then subjects should vote not guilty because the scenario is atypical.

Method

Subjects and procedure. One hundred fifty-two introductory psychology students from Northwestern University participated in partial fulfillment of a course requirement. Subjects listened to an audiotape of the judge’s instructions, then read several short fact situations and made verdict decisions for each. Unlike the previous studies, subjects made decisions for only one crime category, to prevent any carryover of instructions from one category to another. At least 30 subjects judged the scenarios for each crime (31 for burglary and assault); half of the subjects heard preliminary instructions only, and half heard preliminary plus substantive instructions. Subjects completed the questionnaires (described below) at their own pace and were debriefed.

Materials. The two typical and two atypical scenarios for each crime from Experiment 2 were used as stimuli. For each crime, there were also two filler scenarios designed to produce not-guilty verdicts so that there would be a balance of guilty and not-guilty votes for each crime. There were three random orders of scenarios within each crime category. No significant main effects of order were obtained, so the data were collapsed across versions.3

3 Marginally significant main effects of order on the proportion of guilty verdicts were obtained for the assault (p = .09) and kidnapping.
New robbery and murder scenarios. Three new murder scenarios and three new robbery scenarios were written to address the sufficiency issue. These scenarios included very few characteristic features so that they would be perceived by subjects as atypical of the category. Eleven student judges rated these scenarios as quite atypical (M = 2.94 for robbery, and M = 2.91 for murder on a 7-point scale ranging from not at all typical [1] to very typical [7]). Each scenario included one feature that could operate as a sufficient condition for categorization.

“Victim is present” and “something of value is taken” appeared in both the typical and atypical robbery scenarios, and are potentially sufficient conditions for guilt. Three new robbery scenarios tested the sufficiency of these features (see Appendix B for an example): One described an encounter in which the victim was present (but nothing of value was taken), one described a theft in which something of value was taken deliberately (but the victim was not present), and one described an incident in which something of value was taken accidentally (and the victim was not present).

Violence and intentionality were the two features of interest for the murder category. Among the new murder scenarios, one described an intentional (but nonviolent) killing (see Appendix B), and one described a violent (but unintentional) killing. A third scenario described an accidental killing (neither violent nor intentional) to determine whether it is sufficient for guilt that a victim dies at the hands of the defendant. Violence and intentionality were not tested as necessary conditions for guilt in Experiment 1, because they did not meet the 30% frequency criterion adopted in that study. To test the necessity of these features in this experiment, two scenarios were written that (a) included many of the features from subjects’ feature lists so that they would be perceived as typical of the category and (b) omitted one of the potential necessary conditions. Eleven student judges rated these scenarios as quite typical, M = 5.77 on a 7-point scale, ranging from not at all typical (1) to very typical (7). One scenario omitted the feature “intentional,” and the other omitted the feature “violent.” If the omitted feature is necessary for categorization, then subjects should vote not guilty when the feature is not present, even though the scenario is typical in other ways. If the omitted feature is not necessary for guilt, then subjects should vote guilty because the scenario is typical. Altogether, then, there were five new murder scenarios: two testing the necessity of violence and intentionality and three testing the sufficiency of violence, intentionality, and responsibility for the victim’s death.

Results and Discussion

Are the substantive instructions effective? If the substantive instructions (i.e., definitions of the crimes) effectively convey the legally appropriate decision strategy, then instructed subjects’ decisions should no longer be influenced by typicality; they should vote guilty at a high rate for both typical and atypical category members. Table 4 shows the conviction rates of subjects in each instruction condition. As predicted, the pattern of results for subjects who heard the preliminary instructions only was identical to that obtained in Experiment 2 (see the second and third columns of Table 4). Over all five crimes, typical scenarios produced more guilty verdicts than atypical scenarios, t(75) = 7.19, p < .0001. This pattern held for assault, t(14) = 5.14, p = .0002; burglary, t(15) = 8.59, p < .0001; and kidnapping, t(14) = 4.18, p = .0009; with robbery and murder producing uniformly high conviction rates. Subjects who heard the preliminary instructions only were not expected to change their decision strategies, because these instructions provide no strategy-relevant information. It was possible that the conviction rates for both typical and atypical scenarios would be lower in Experiment 3 than they were in Experiment 2 because the preliminary instructions specify a relatively high decision threshold. However, the conviction rates are similar in magnitude to those obtained in Experiment 2 (see Table 3).

Did subjects who heard both the preliminary instructions and the definition of the crime charged make accurate decisions? As shown in columns 4 and 5 of Table 4, the pattern of convictions for these subjects was the same as for the preliminary instructions-only subjects. Over all five crimes, typical scenarios produced more guilty verdicts than atypical scenarios, t(75) = 6.23, p < .0001. This pattern held for assault, t(15) = 3.87, p < .002; burglary, t(14) = 5.39, p = .0001; and kidnapping, t(14) = 3.16, p < .007; with similarly high conviction rates on the typical and atypical robbery and murder scenarios, both ts(14) = 1.00, ns. A two-way ANOVA was performed on the overall data (collapsed across the five crimes), with instruction as a between-subjects factor and typicality as a within-subjects factor. This analysis yielded a main effect of typicality, F(1, 150) = 90.57, p < .0001, no main effect of instruction, F(1, 150) = 0.46, ns, and no interaction, F(1, 150) = 1.42, ns. Separate ANOVAs for each crime revealed main effects of instruction and no interactions for any of the crimes. These results indicate that hearing the judge’s instructions on the crime definitions did not alter the strategies subjects used to make their verdict decisions. Typicality still influenced categorization for assault, burglary, and kidnapping (ps < .0001) but not for murder or robbery. There were no main effects of instruction for assault, burglary, and kidnapping, even when subjects were instructed in the legally correct decision strategy.

Decision processes for murder and robbery. The proportion

### Table 4

<table>
<thead>
<tr>
<th>Crime</th>
<th>Preliminary instructions only</th>
<th>Preliminary instructions plus crime definition</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Typical</td>
<td>Atypical</td>
</tr>
<tr>
<td>Assault*</td>
<td>.87</td>
<td>.40</td>
</tr>
<tr>
<td>Burglary*</td>
<td>1.00</td>
<td>.22</td>
</tr>
<tr>
<td>Kidnapping*</td>
<td>1.00</td>
<td>.50</td>
</tr>
<tr>
<td>Murder</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Robbery</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Overall*</td>
<td>.97</td>
<td>.62</td>
</tr>
</tbody>
</table>

* Main effect of typicality in two-way analysis of variance. There were no main effects of instruction and no interactions.

(p = .06) scenarios. For assault, typicality interacted significantly with order (p = .02). For all three orders, typical scenarios produced a higher conviction rate than atypical scenarios (as predicted), but the difference was larger for Order 1 than for the other orders. For kidnapping, typicality interacted marginally with order (p = .06). Again, for each order, typical scenarios produced more convictions than atypical scenarios, but the difference was smaller for Order 2 than the others. No other interactions approached significance.
of guilty votes for each of the new robbery and murder scenarios was computed for the preliminary-instructions-only subjects to determine the sufficiency (and necessity) of the features of interest for naive decision makers. The scenarios testing sufficiency were atypical of the category. Thus, if a target feature was not sufficient for guilt, then the conviction rate for that scenario should be low. However, if a target feature was sufficient, then the conviction rate should approach 100%. The proportion of guilty votes for each of these scenarios was tested against 100% to determine the sufficiency of the target features. 

Two of the target features for the robbery category produced conviction rates far below the 100% criterion, indicating that neither was a sufficient condition for guilt. The victim present scenario yielded a 20% conviction rate ($z = -30.8, p < .0001$), and the accidental taking scenario produced a 40% conviction rate ($z = -23.1, p < .0001$). The third target feature, deliberate taking, produced an 87% conviction rate. This was significantly lower than the 100% sufficiency criterion ($z = -5.0, p < .0001$) but significantly higher than the 50% one would expect if subjects were simply guessing ($z = 2.33, p < .01$, one-tailed). To verify that it was the deliberate taking feature of this scenario that produced the high conviction rate, and not some other feature or peculiarity of the fact situation, 19 additional subjects read this scenario with the deliberate taking feature removed. This omission dropped the conviction rate from 87% to only 5%. Apparently, this feature carries considerable weight for verdict decisions, although it is not strictly sufficient; other features still play some relatively small role in the decision.

Note that the deliberate taking scenario (see Appendix B) did not meet the legal requirements for robbery. The defendant took an object but did not take it from the victim by force or threat of force. Therefore, the defendant's crime was not a member of the robbery category, and guilty votes on this scenario represent false alarm errors. The 87% error rate obtained for this fact situation indicates that subjects did not know and use the correct defining features of the robbery category. Had they done so, they would have correctly rejected this scenario as an instance of robbery. This finding suggests that subjects' correct categorizations of the atypical robbery scenarios were probably not due to knowledge of the defining features of robbery but were rather a product of including the deliberate taking feature in these scenarios.

For the murder category, two of the target features produced conviction rates far below the 100% criterion, indicating that neither was sufficient for guilt. The violent killing scenario produced no guilty verdicts ($z = -38.5, p < .0001$), and the accidental killing scenario produced only a 7% conviction rate ($z = -35.8, p < .0001$). The third target feature, intentional killing, produced a 67% conviction rate, which was significantly lower than 100% ($z = -12.7, p < .0001$) but marginally higher than the 50% expected if subjects were simply guessing ($z = 1.32, p = .09$, one-tailed). To verify that intentionality was producing this moderately high conviction rate, 19 additional subjects read this scenario with the intentional killing feature removed (the fifth sentence of the scenario was omitted; see Appendix B). With this omission, the conviction rate dropped to 26%, suggesting that intentionality is a relatively important feature for murder verdicts, even though it is not sufficient.

Are violence and intentionality necessary for the categorization of murders? The scenarios testing necessity were typical of the category and omitted the potentially necessary condition. With these scenarios, then, high conviction rates mean that the omitted feature is not necessary but is merely associated with the category. The scenario testing the necessity of violence produced a 100% conviction rate ($z = 38.5, p < .0001$), and the scenario testing intentionality produced an 80% conviction rate ($z = 30.8, p < .0001$). Both conviction rates significantly exceeded the 0% performance criterion (and a more conservative 50% criterion), indicating that neither feature was necessary for guilt. It is important to note that intentionality, or knowledge of consequences, is a necessary condition for guilt under the law (see Appendix A). The fact that subjects did not treat it as a necessary condition suggests that they did not know the correct defining features of the murder category.

Apparently, subjects tend to use the terms robbery and murder as general labels for deliberate takings and intentional killings, respectively. Under the law, many deliberate takings are not robbery, but theft, and many intentional killings are not murder, but manslaughter. However, it appears that these distinctions are often lost on legally untrained decision makers. It is not surprising, perhaps, that robbery is used as a general term for deliberate taking. Television and movies subsume true robberies, thefts, and burglaries under the label robbery. Burglary victims, for example, often exclaim "We've been robbed!" Does this mean that subjects' prototypes for robbery and murder do not influence their categorization decisions? Experiments 2 and 3 demonstrated that the accuracy of subjects' decisions was unaffected by typicality, but it is possible that people's prototypes influenced the ease of categorization. If typical category members are easier to identify than atypical category members, then subjects should be able to make their verdict decisions more quickly for typical than atypical fact situations.

To test this possibility, 19 subjects made verdict decisions for the robbery and murder scenarios presented on a computer that recorded decision times. Subjects made decisions on 16 scenarios: 2 practice crimes at the beginning of the session (1 burglary and 1 assault) to familiarize subjects with the task, the 2 typical and 2 atypical robberies, the 2 typical and 2 atypical murders, and 6 filler crimes (3 robberies and 3 murders) that were not category members and were included to provide a balance of guilty and not-guilty votes. Each fact situation was presented to subjects on two or three successive screens, and subjects pressed the space bar to scroll through the fact screens. After each scenario, subjects pressed the space bar to advance to a decision screen, which asked, "Is (defendant) guilty of (crime)?" The name of the particular defendant and the particular crime were inserted for each scenario. To respond to this question, subjects pressed a key marked yes or a key marked no, which were adjacent to each other in the middle of the keyboard. Subjects were not given any instructions about which hand to use when responding, so they could use whichever hand they preferred.

---

4 A $z$ test for binomial proportions cannot tolerate a 100% performance criterion because the standard error of the mean would then be zero and $z$ would be undefined. To avoid this problem, a number close to 1.00 (99) was used as the criterion.
hand was comfortable. Response time was recorded by the computer from the time the decision screen was displayed to the time subjects pressed the yes or no key. Subjects were told that their decisions would be timed and that they should respond as quickly as possible without making mistakes. There were 10 random orders of scenarios used in this experiment, which met the following randomization constraints: (a) every order began with the 2 practice crimes, (b) the first half of the scenarios (after the practice) included 1 typical and 1 atypical murder, 1 typical and one atypical robbery, and (at least) 1 filler murder and 1 filler robbery, and (c) no more than 2 crimes in succession were from the same category (i.e., robbery or murder).

Consistent with the previous experiments, the typical and atypical robbery scenarios produced uniformly high conviction rates (.95 and .97, respectively; z = -0.31, ns), as did the typical and atypical murder scenarios (.95 and .92, respectively; z = 0.38, ns). The issue of interest was whether it is easier, and therefore quicker, for subjects to identify typical than atypical category members. Decisions that a target is not a category member (not-guilty responses) are thus not relevant and were excluded from the analyses. This involved three responses for robbery and five for murder. One subject voted not guilty for three out of four critical murder scenarios; that subject's data had to be excluded from the murder analyses because a within-subjects comparison of response times on typical and atypical scenarios could not be made. This meant that 19 subjects provided data for the robbery analyses, and 18 subjects provided data for the murder analyses.

Before the decision times for typical and atypical scenarios were compared, outlying responses were removed. Responses that were more than 2 standard deviations from the mean decision time for a category were considered outliers. This resulted in the exclusion of four responses in the robbery category and two responses in the murder category. Because each subject made decisions on two typical and two atypical scenarios in each category, the removal of outliers meant that no subjects were excluded from the analyses, only a small number of scenario responses. For each subject, the decision times for the two typical and two atypical scenarios of each crime type were averaged, and this mean response time was used in the analyses. When one of the two response times was an outlier or a not-guilty decision, it was excluded, and the response time for the other scenario was used in the analyses.

A matched-pairs t test compared the decision times for the typical and atypical scenarios of each crime category. For robbery, typical scenarios (M = 3.52 s) were identified no more quickly than atypical scenarios, (M = 3.38 s; t(18) = -0.51, ns). Apparently, subjects' prototypes did not affect the ease of categorizing the robbery scenarios. For murder, however, typical scenarios (M = 2.62 s) were identified more quickly than atypical scenarios, (M = 3.29 s; t(17) = 2.10, p < .05, one-tailed), indicating that it was easier for subjects to identify the typical scenarios as murders (i.e., as category members) than to identify the atypical scenarios.

General Discussion

The purpose of this research was to explore the role of the law in the verdict selection process. How do people use legal concepts to choose a verdict? There are two potential sources of information on the law that are available to jurors: the judge's instructions and their own prior knowledge. Jurors are supposed to rely solely on the judge's instructions, but if they have constructed naive representations of crime categories before coming to trial, these concepts may interfere with their understanding and use of the judge's instructions. The starting point for this research, then, was to discover whether laypeople have constructed representations of crime categories and, if so, to investigate the organizational structure of those representations.

The results of Experiment 1 demonstrated that people have indeed formed naive concepts of crime categories. The features of their representations did not correspond to the correct defining features of the crimes, suggesting that subjects' prior knowledge is a potential source of conflict with the judge's instructions. Subjects' representations appeared to contain category prototypes that influenced their perceptions of fact situations; some category members were perceived as more typical than others. Experiment 2 investigated the impact of these prototypes on categorization. In Part 1, subjects provided the correct category label more often for typical than atypical category members for assault, burglary, kidnapping, and robbery but not for murder. In Part 2, subjects voted guilty more often for typical than atypical category members for assault, burglary, and kidnapping but not for murder or robbery. In these experiments, then, uninstructed decision makers used their prototypes as guides for verdict selection when deciding assault, burglary, and kidnapping cases.

The response time data revealed that subjects' prototypes did play a role in decision making for murders. Typical category members were identified more quickly than atypical category members, even though accuracy was high for both types of scenarios. In practice, of course, it matters little how quickly jurors are able to categorize fact situations, particularly when the difference is of the order of half a second. Nevertheless, these data do indicate that subjects' prototypes are active and involved in the decision process for the murder category. It appears that for robbery, subjects were most influenced by the deliberate taking feature. Both the typical and atypical robberies involved deliberate takings, and this may have accounted for the high accuracy rates and comparable response times for these scenarios. In fact, this feature was so influential that its inclusion in a nonrobbery scenario resulted in 87% false alarms.

Before hearing the judge's instructions, the only source of information people have for decision making is their own prior knowledge. They may rely on their prototypes because they have no other decision strategy available to them. However, Experiment 3 revealed that the verdict decisions of instructed subjects did not differ from those of uninstructed subjects; instructed decision makers did not disregard their prior knowledge in favor of the defining features presented by the judge. The judge's instructions for these crimes differ considerably in complexity (see Appendix A). Assault requires the proof of only one proposition; burglary requires proof of three. Kidnapping is more complex, with several conditional propositions. However, the judge's instructions did not change subjects' decision strategies for any of these crimes. In the simplest case of assault,
typicality still influenced verdict selection. Therefore, even relatively simple instructions failed to revise subjects’ existing concepts enough to improve the accuracy of their decision making.

Implications for Jury Instruction

These results provide strong evidence that laypeople are not blank slates with regard to the law; their prototypes of crime categories can influence both their perceptions of fact situations and their categorization decisions. The objective of jury instruction, then, must be concept revision, not merely concept formation. To be effective, the judge’s instructions must accomplish several goals. First, they must inform jurors that their preferred decision strategy is inappropriate, namely, that the typicality of the crime is irrelevant. Second, they must explain the proper strategy, namely, that verdict decisions must be based on the defining criteria specified by the law. Third, the instructions must make clear what are the necessary and sufficient conditions for guilt. At present, it is unclear where the current instruction process breaks down. We know only that the instructions are not effective in conveying the appropriate decision strategy. Future research should investigate methods of presenting the instructions that successfully revise both the content of jurors’ existing concepts and the way they use that content for verdict selection.

These results also have interesting implications for proposed linguistic and procedural reforms of jury instruction. Some researchers have advocated rewriting jury instructions in simpler language to improve jurors’ comprehension of the law (Charrow & Charrow, 1979; Elwork, Sales, & Alfini, 1977; Severance & Loftus, 1982). The present research reveals a potential cost of such simplification. Colloquial terms carry colloquial baggage, some possibly correct, some incorrect. Wholesale replacement of legal terms with simple language may activate a host of associated concepts that are useful for everyday decision making but are legally incorrect or irrelevant. Rewriting must be undertaken only when jurors’ existing knowledge structures will help them to understand and use the information presented at trial and must be avoided when those knowledge structures contain incorrect information that might impair the accuracy of jurors’ decision making. Only careful testing of individual instructions can distinguish these situations.

Much has also been made of the potential for pretrial instruction to improve jurors’ decisions (Goldberg, 1981; Kassin & Wrightsman, 1979; Prettyman, 1960; V. L. Smith, 1991). The present experiments suggest that it is not that simple. In Experiments 2 and 3, instructed subjects heard the law, then read the fact situations and made their verdict decisions. These subjects were preinstructed on the law, but their verdict decisions were no more accurate than those of uninstructed subjects. These findings underscore the importance of the conflict between people’s prior knowledge of the law and the judge’s instructions. If a reform of the instruction process fails to revise jurors’ existing concepts, it is unlikely to have salutary effects on decision accuracy. Simply presenting the law before the facts apparently does not achieve this goal of concept revision.

External Validity

Generalizations of these findings must, of course, be made cautiously. To investigate jurors’ prior knowledge of the law in a controlled fashion, certain compromises on external validity were necessary. Consideration of these compromises suggests some important directions for future research. For example, subjects in these experiments made decisions on very limited evidence, responding to fact situations that were, on the average, only six sentences long. This is quite different from the usually large, complicated, and often conflicting body of evidence presented in actual trials. Because this research explored the role of subjects’ prior knowledge of the law in the verdict selection process, it was important that all subjects based their decisions on the same evidence. To this end, the facts were presented in a straightforward, coherent, and internally consistent manner: subjects did not have to interpret or make inferences about the evidence to generate a plausible representation of the facts. This simplification was necessary to minimize variability in verdict choices due to differing interpretations of the evidence. Having found that people’s prior knowledge of legal concepts can have powerful effects on their verdict choices, future research should replicate these effects using more realistic stimulus trials.

Another departure from actual jury trials was the subject population used in these experiments. All subjects were undergraduates at Northwestern University, who are younger and better educated than the average juror. It is possible that on these tasks the performance of actual jurors might differ from the performance of students. The data reported in this article revealed that subjects had constructed prototype representations of the target crimes (the same way they structure other categories), that these naive representations influenced decision making, and that subjects’ strategy for using these representations was resistant to change by instruction. If students and community jurors differ, one might expect the students to perform more accurately on these tasks than the average juror. Students are, perhaps, more likely to have discovered or learned in the course of their studies that legal decision making differs from other kinds of categorization tasks, are more likely to have learned how to make decisions from a set of necessary and sufficient conditions, and generally are more practiced at acquiring and using new decision strategies. To the extent that students in these experiments were unable to perform these tasks, it is unlikely that a more diverse population of jurors would be better informed or strategically more facile. However, future research should test this prediction using a more diverse subject sample.

Implications for Decision Making

This research addressed the generality of prototype-based decision making in a decision context that required a competing strategy. Previous research on social categorization involved decisions or categories that were not well-defined. In the legal context, the boundaries of crime categories are relatively clearly specified, and there is a normatively appropriate decision strategy. However, in this decision context as well, it appears that people have constructed prototypes of the relevant categories and have used them as guides for decision making. It is perhaps
not surprising that naive decision makers are unaware of the legal constraints on verdict decisions. It might be difficult to discover from informal sources that verdict decisions should be based on a set of defined features. However, even when subjects were instructed on the correct decision process, their prior knowledge still influenced their decisions.

These findings are consistent with research in other areas of social psychology demonstrating the potential dangers of prior knowledge or theories for accurate judgment and decision making. For example, research on expectancy effects has revealed that having expectations about others can influence our impressions of them and cause us to elicit belief-confirming behaviors from them (Kelley, 1950; Rosenthal & Jacobson, 1968; Snyder, Tanke, & Berscheid, 1977). In a similar vein, our prior theories about a decision domain can lead us to substantially overestimate the true correlations among variables (Chapman & Chapman, 1967; Jennings, Amabile, & Ross, 1980) or to reject diagnostic information that is inconsistent with our beliefs (Lord, Ross, & Lepper, 1979; Ross, Lepper, & Mackie, 1984). A prototype analysis of psychological situations (Cognitive Psychology, 14, 45-77).


References


Appendix A

**Illinois Pattern Jury Instructions (1981)**

**Section 11.02 Assault:** To sustain the charge of assault, the State must prove the following proposition: That the defendant engaged in conduct which placed the victim in reasonable apprehension of receiving bodily harm. If you find from your consideration of all the evidence that this proposition has been proved beyond a reasonable doubt, you should find the defendant guilty. If you find from your consideration of all the evidence that this proposition has not been proved beyond a reasonable doubt, you should find the defendant not guilty.

**Section 11.06 Burglary:** To sustain the charge of burglary, the State must prove the following propositions: First, that the defendant knowingly entered a building; and Second, that the defendant did so without authority; and Third, that the defendant did so with the intent to commit a felony. If you find from your consideration of all the evidence that each one of these propositions has been proved beyond a reasonable doubt, you should find the defendant guilty. If you find from your consideration of all the evidence that any one of these propositions has not been proved beyond a reasonable doubt, you should find the defendant not guilty.

**Section 14.04 Kidnapping:** To sustain the charge of kidnapping, the State must prove the following propositions: First, that the defendant acted knowingly; and Second, that the defendant secretly confined the victim against his will; and Second, that the defendant, by force or threat of imminent force, carried the victim from one place to another place; and Third, that when the defendant did so he intended secretly to confine the victim against his will. If you find from your consideration of all the evidence that any one of these propositions has been proved beyond a reasonable doubt, you should find the defendant guilty. If you find from your consideration of all the evidence that any one of these propositions has not been proved beyond a reasonable doubt, you should find the defendant not guilty.

**Section 17.06 Robbery:** To sustain the charge of robbery, the State must prove the following propositions: First, that the defendant took property from the person or presence of the victim; and Second, that the defendant did so 1) he intended to kill or do great bodily harm to the victim; or 2) he knew that his act would cause death or great bodily harm to the victim; or 3) he knew that his acts created a strong probability of death or great bodily harm to the victim. If you find from your consideration of all the evidence that any one of these propositions has been proved beyond a reasonable doubt, you should find the defendant guilty. If you find from your consideration of all the evidence that any one of these propositions has not been proved beyond a reasonable doubt, you should find the defendant not guilty.
**Appendix B**

Examples of Crime Scenarios

**Kidnapping**

Many characteristic features:

Tony was playing ball with his friends in the playground of his elementary school one afternoon. Ken knew that Tony's parents were very wealthy and very protective of their child. Ken called to Tony and waved him over to his car. When Tony came over, Ken asked if he wanted to go for ice cream. Tony said, "Sure," and got in the car. Ken bought Tony an ice cream, then took him to a motel room, where he tied him to a chair, gagged him, and told him if he made any noise he'd kill him. Ken then took a note demanding $500,000 for Tony's safe return, and left it in the mailbox of Tony's parents' house.

Intermediate number of characteristic features:

Ron and Helen had always wanted children, but were unable to have any. They had also been unsuccessful in their attempts to adopt, so they decided to get a child themselves. One day, Ron spotted a woman in a book store with a toddler. The woman paid little attention to the child, and let him wander among the books without supervision. When the child walked to the back of the store, out of his mother's sight, Ron picked him up and left the store through an exit at the back. He took the child home, planning to tell his neighbors that he came from the adoption agency.

Few characteristic features:

Leon was an investigator with the Drug Enforcement Administration. He discovered the headquarters of a major drug ring, and went there to gather evidence against them. He hid behind a stack of boxes and listened to their conversation. In their discussion, the dealers revealed everything Leon needed to bust them. When he tried to sneak out, he knocked over one of the boxes and was discovered. The dealers told Leon to tell them everything he knew or they would beat him until he did. Leon wouldn't tell them anything, so the dealers locked him in their hide-out and left him there.

**Assault**

Typical assault:

Tim was walking down a dark street one night when he saw his girlfriend talking to another man. Tim got angry, pulled a knife from his pocket and walked up to the man. He held up the knife menacingly, and said, "Get lost." The man was afraid Tim would use the knife, so he put up his hands and backed away. Tim walked toward him, jabbed the knife in his direction and said, "Come near her again and I'll kill you." The man nodded and Tim and his girlfriend pushed past him.

Atypical assault:

When Frank applied for a hunting license, the game warden asked to inspect Frank's firearms. He found that Frank's shotgun was faulty and said he had to confiscate it. Frank got angry and said, "Nobody's gonna take my gun. Now hand it back, or I'll have to take it back myself, and you wouldn't want that. You have such a nice face; don't make me ruin it." Frank had quite a reputation as a fighter, and the warden was a much smaller man than Frank. Afraid Frank would injure him, the warden handed the gun back and decided to let the sheriff's office handle the confiscation.

**Robbery**

Typical robbery:

Bob needed cash to pay off a bet. The Martins, an old couple who lived down the street, always had a lot of money in their apartment, and were an easy mark. One night, Bob saw Mr. Martin walking home, and followed him. When Mr. Martin opened the door to his apartment, Bob pulled out his gun, walked up behind him and pushed him inside. Bob told him to hand over all his money. Mr. Martin was afraid Bob was going to shoot him, so he gave Bob $200 in cash.

Atypical robbery:

Andy was standing on a street corner one day when a BMW stopped in front of him at a stop light. The window on the passenger's side was rolled down, and he could see a woman's purse on the floor. Andy quickly opened the passenger door, stuck his head in the car and said, "Don't move and I won't hurt you." The driver just nodded. Andy grabbed the purse and ran away.

**Sufficiency of deliberate taking:**

Roger was in a park one Saturday afternoon when he noticed a portable cassette recorder propped up against a tree. Roger looked around, but there was no one in sight. He sat down on a bench and waited for several minutes, but no one appeared. Finally, he walked over and picked up the cassette recorder. He walked home with it as fast as he could.

**Murder**

Sufficiency of intentional killing:

Ben and his business partner, Terry, were standing on a train platform when Terry told Ben he had used company money to pay his gambling debts, and they were bankrupt. Ben had invested everything in the business. Enraged, he grabbed Terry and held him over the edge of the platform. A train was rapidly approaching the station, and Terry was afraid Ben would let go. Ben knew Terry had a heart condition, and that stress would kill him, but he didn't care. Ben pulled him back onto the platform just as the train roared by. Terry clutched his chest and died of a heart attack.

Received February 27, 1990
Revision received January 29, 1991
Accepted June 26, 1991
When Prior Knowledge and Law Collide

Helping Jurors Use the Law*

Vicki L. Smith

Jurors are supposed to rely on the judge’s instructions for verdict selection. However, recent research indicates that people have constructed naive representations of crimes that conflict with the judge’s instructions and that influence decision making. The present research explored potential solutions to this conflict. Two experiments revealed that the problem cannot be circumvented by avoiding people’s prior knowledge; subjects activated and used their prior knowledge of crimes even when the crime name was withheld. Experiment 3 demonstrated that a supplementary instruction to disregard prior knowledge was also ineffective. Experiment 4 revealed that a supplementary instruction designed to revise subjects’ existing representations did improve decision accuracy. These experiments indicate that the conflict between people’s prior knowledge and the law cannot easily be avoided or disregarded, but its impact can be reduced by revising people’s existing concepts.

A juror’s duty is to make sense of the evidence presented at trial, to listen to the judge’s instructions on the law, and then to integrate facts and law into a legally appropriate verdict. There are two sources of information jurors may use for fact-finding: (1) the evidence and testimony presented at trial and (2) their prior knowledge of the physical and social world. The latter may be used for determining the credibility of the witnesses, filling gaps in the testimony, and resolving disputed factual issues. This use of jurors’ prior knowledge is encouraged in the judge’s instructions to the jury: “In determining the credibility of a witness you may consider any matter that has a tendency in reason to prove or disprove the truthfulness of his (sic) testimony.” (California Jury Instructions—Criminal, § 2.20, 1970); “You should consider all the evidence in the light of your own ob-

* I would like to thank Norma Fiedotin, Shana Fineburg, and Chinazo Opia, for their help in testing subjects. Correspondence concerning this article should be addressed to Vicki L. Smith, Department of Psychology, 102 Swift Hall, Northwestern University, Evanston, IL 60208.
servations and experience in life." (Illinois Pattern Jury Instructions, Criminal, § 1.01, 1981). In contrast to the evaluative freedom jurors are permitted as fact-finders, their use of the law is highly constrained. Jurors are told that they must follow the law presented in the judge’s instructions: “It is my duty to instruct you in the law that applies to this case and you must follow the law as I state it to you.” (California Jury Instructions—Criminal, § 1.00, 1970). Jurors are assumed either to have no prior knowledge of the law or to disregard that prior knowledge and rely on the judge’s instructions when making their verdict decisions.

Although most jurors have had no formal training in law or legal reasoning (American Bar Association, 1968; Loh, 1984), it is likely that they have had some exposure to the law through informal sources, such as movies, television, newspapers, and word of mouth. From these sources, jurors may have abstracted representations of many legal concepts that they can access for decision making, concepts such as the presumption of innocence, the reasonable doubt standard, or what constitutes crimes like burglary, assault, and murder. Rather than a blank slate, then, jurors may bring to trial naive representations of legal concepts, and this prior knowledge may serve as a source of information about the law. Recent research has revealed that people have indeed constructed naive representations of crime categories and, as described shortly, this information influences people’s perceptions of fact situations and their verdict decisions (Smith, 1991b). Furthermore, this prior knowledge continues to be influential even when people hear the judge’s instructions on the law. The goal of the present research was to explore ways of reducing or eliminating the influence of jurors’ prior knowledge of the law so that they rely on the judge’s instructions for verdict selection.

The Nature of Jurors’ Prior Knowledge

To determine whether people have constructed naive representations of crime categories, Smith (1991b) asked subjects to list the features they believed characterize crimes like assault, kidnapping, and robbery. Subjects were readily able to do this, listing the attributes of the victim and the perpetrator, and the actions, intentions, and motives of the criminal. The features listed most frequently by subjects tended to be legally incorrect, irrelevant, or incomplete. For example, subjects typically described robbery as (a) the taking of money or other valuables, (b) from a home, (c) by an armed perpetrator. In contrast, the law defines robbery as the taking of property from a victim by force or threat of force. So, subjects were correct that robbery involves “taking,” but, under the law, the item taken need not be valuable, the location need not be a home, and the perpetrator need not be armed. In addition, the law requires that the victim be present when the object is taken and that the perpetrator use force or threat of force. In short,

1 Some states (e.g., Maryland and Indiana) instruct jurors that they have the right to nullify the law and vote not guilty if they believe a guilty verdict would be unjust under the circumstances (Hans & Vidmar, 1986). This instruction is a rare exception to the expectation that jurors will base their verdicts on the law as presented by the judge.
subjects’ feature lists indicated that they had constructed representations of robbery and other crime categories that they could access when needed: however, much of the content of those representations was legally incorrect and is thus a potential source of conflict with the judge’s instructions.

Lay wisdom and the law also appeared to diverge with respect to decision strategy (Smith, 1991b). Under the law, crime categories (like assault and burglary) are classically defined by a set of singly necessary and jointly sufficient features: If all the necessary conditions are met beyond a reasonable doubt, the juror is supposed to vote guilty; if any necessary condition is not met beyond a reasonable doubt, the juror is supposed to vote not guilty. In contrast, it appears that the features contained in people's naive representations of crime categories do not operate as necessary conditions. Rather, these naive concepts appear to contain prototypes (or typical exemplars) of crime categories. Prototypes are summary representations of the characteristics of category members. These representations exhibit a graded structure; category members possessing many characteristic features of the prototype are perceived to be more typical of the category than members possessing few characteristic features (see Smith & Medin, 1981). The perceived typicality of a target can influence both the speed and accuracy of categorization (Rosch, 1975; Rosch & Mervis, 1975).

To determine whether people’s representations of crime categories have a graded structure, Smith (1991b) had subjects read several short scenarios, each describing the facts of a criminal encounter, and rate how typical each scenario was of the crime in question. Some of the scenarios contained many of the characteristic features of the crime from subjects’ feature lists, other scenarios contained few characteristic features. All of the scenarios met the legal requirements for the particular crime, so all were category members. For each crime (e.g., burglary, assault), subjects’ judgments did indeed exhibit the graded structure of a prototype representation; scenarios containing many characteristic features were perceived to be more typical of the crime category than scenarios containing few such features. Furthermore, when asked to choose verdicts for these scenarios, subjects voted guilty significantly more often for the typical than the atypical crimes, indicating that this graded structure also influenced categorization decisions.

Together, these findings suggest that people are not blank slates with regard to the law. They have constructed naive representations of crime categories that influence their perceptions of fact situations and their verdict preferences. There appear to be two important ways in which people’s prior knowledge of the law conflicts with proper legal decision making. First, the content of people’s representations is largely erroneous; features people believe to be characteristic of a crime are often irrelevant under the law. Second, people’s naive decision strategy is legally inappropriate; verdict decisions are influenced by people’s prototypes of crime categories when they should be based on a set of specific necessary and sufficient conditions. To be effective, then, it appears that the judge’s instructions must correct both the content of people’s naive representations and their strategy for using that content.
The Judge's Instructions

Substantive jury instructions define the crime charged against the defendant; they specify the legally appropriate decision criteria and explain how to use those criteria for verdict selection. This means that jurors are given all the information they need to make legally correct verdict decisions. However, experiments testing the effectiveness of jury instructions for educating jurors about the law have revealed that mock jurors' comprehension of the instructions is quite poor. In many of these experiments, instructed jurors have performed (a) no better than uninstructed jurors (Elwork, Sales, & Alfini, 1977; Severance & Loftus, 1982; Smith 1991a, 1991b), and (b) at chance levels on tests of comprehension for the instructions (Elwork et al., 1977; Smith, 1991a). Smith (1991b) compared the verdicts of instructed and uninstructed mock jurors to determine whether the substantive jury instructions effectively educated jurors about proper legal decision making. If so, instructed subjects' verdict choices should no longer be influenced by the typicality of the fact situation; they should vote guilty at a high rate for both typical and atypical category members. In fact, however, the verdicts of instructed and uninstructed subjects did not differ; both groups voted guilty more often for typical than atypical crimes. Apparently, then, the instructions did not discourage subjects from using their prior knowledge for decision making.

Some research has demonstrated benefits of jury instruction. In some experiments, instructed subjects show better comprehension of the law than uninstructed subjects (Buchanan, Pryor, Taylor, & Strawn, 1978; Strawn & Buchanan, 1976). These differences are statistically significant, although they tend to be small. Other experiments have shown that subjects' decisions can be altered with certain variations in the requirements of proof instructions (Kagehiro & Stanton, 1985; Kerr et al., 1976). This research indicates that, at least in some cases, subjects are willing and able to use judge's instructions for decision making. Apparently, then, the problem is not that people are insensitive to instruction, but that they are unwilling or unable to use many of the instructions they hear.

To improve jurors' understanding of the law, several researchers have proposed rewriting jury instructions in simpler language, and they have demonstrated improved comprehension of rewritten instructions (Charrow & Charrow, 1979; Elwork et al., 1977; Severance & Loftus, 1982). Other researchers and legal scholars have advocated procedural changes in the presentation of instructions to improve jurors' comprehension of the law. Jury instructions are usually delivered orally by the judge after the evidence has been presented and just before deliberation. Reformers have proposed that presenting the instructions before trial, allowing jurors to take notes on the instructions, or giving them written copies of the instructions might help them learn the law (Goldberg, 1981; Prettyman, 1960; Sand & Reiss, 1985). There is some evidence that presenting the instructions both before and after trial improves jurors' comprehension (Smith, 1991a), but note-taking and providing written copies of the instructions have been disappointingly ineffective (Hastie, 1982; Heuer & Penrod, 1988, 1989).

Although rewriting and pretrial instruction have produced some improvements, they have not solved the comprehension problem; there are still important
gaps in jurors’ understanding even when these procedures are implemented. This suggests that other obstacles to comprehension are operating. People’s prior knowledge of the law may be one such obstacle. Effective jury instruction may require greater attention to people’s naïve representations of legal concepts; jurors’ prior knowledge of the law may interfere with their understanding and use of the judge’s instructions. Thus, accurate decision making may depend on resolving the conflict between lay wisdom and law. The goal of the present research was to explore potential solutions to this conflict.

One possibility, tested in Experiments 1 and 2, is to circumvent the conflict by not activating jurors’ prior knowledge of the law. If jurors do not access their naïve concepts, then the judge’s instructions are the only source of information about the law available for decision making. Jurors would have to rely on the instructions for choosing a verdict because there is no competing information. Another possible solution, tested in Experiment 3, is to inform jurors directly that whatever knowledge of the law they brought with them to trial is irrelevant and must be disregarded for decision making. Such an instruction has two potential advantages. First, it alerts jurors to the possibility that their prior knowledge may be inconsistent with the law, something that is not explicit in current jury instructions, and second, it highlights the importance of the instructions as the basis for decision making. This kind of supplementary instruction may discourage jurors from making assumptions about the law and encourage them to focus greater attention on the decision criteria contained in the substantive definition of the crime charged. A more comprehensive, but more intrusive, potential solution to the conflict is to confront jurors’ misconceptions on a feature-by-feature basis. It may be necessary to give jurors detailed information about which features of their naïve representations should be revised, and how. The substantive jury instructions could then fill the gaps in jurors’ existing representations. Experiment 4 tested the effectiveness of a supplementary instruction that provided a feature-by-feature attack on jurors’ mistaken prior beliefs.

EXPERIMENT 1

Experiment 1 investigated the possibility that the conflict between jurors’ prior knowledge and the law can be avoided. If jurors’ prior knowledge is not activated, they would have to rely on the judge’s instructions for verdict selection because there is no other information available. Current trial procedures make accessing prior knowledge relatively simple; jurors are informed at the start of the trial what crime is charged against the defendant, and this category label can then serve as a retrieval cue for the characteristic features in people’s naïve representations of that crime. If this retrieval cue were not provided, jurors might not be able to activate their prior knowledge of the relevant category and there would be no conflict with the law.

In this experiment, the name of the crime category was withheld in an attempt to prevent jurors from accessing their prior knowledge of burglary. Subjects were
simply told that the defendants were charged with a serious crime. If this intervention is successful, subjects' verdict choices should be unaffected by the perceived typicality of the fact situations. If they do not access the prior knowledge on which typicality judgments are based, there should be no greater tendency to vote guilty for typical than atypical category members. Furthermore, if they understand and use the judge's instructions, the conviction rate for all category members should be near 100%.

Method

Subjects and Procedure

Eighty-five introductory psychology students participated in partial fulfillment of a course requirement. Subjects were randomly assigned to one of four instruction conditions. They listened to an audiotape of the judge's instructions, then chose verdicts for 12 crime scenarios. At the conclusion of the experimental session, subjects were fully debriefed.

Instructions

In the preliminary only instruction condition, subjects were told that the defendants were charged with burglary, then they heard preliminary jury instructions that warned against being influenced by sympathy or prejudice and explained that each defendant is presumed innocent unless the evidence establishes guilt beyond a reasonable doubt (Illinois Pattern Jury Instructions, Criminal, 1981). These instructions are delivered in all criminal trials and contain no information about the substantive legal definition of burglary. So, subjects in the preliminary only condition knew that the defendants were charged with burglary, but they were not told what constitutes burglary under the law. As a result, these subjects had only their prior knowledge of burglary under the law. As a result, these subjects had only their prior knowledge of burglary to guide decision making and they were expected to vote guilty more often for typical than atypical burglary scenarios.

Subjects in the preliminary plus burglary condition were told that the defendants were charged with burglary, then they heard both the preliminary instructions and the legal definition of burglary:

A person commits the offense of burglary when he, without authority, knowingly enters a building with intent to commit a felony therein. To sustain the charge of burglary, the State must prove the following propositions: first, that the defendant knowingly entered a building; and second, that the defendant did so without authority; and third, that the defendant did so with the intent to commit a felony. If you find from your consideration of all the evidence that each one of these propositions has been proved beyond a reasonable doubt, you should find the defendant guilty. If you find from your consideration of all the evidence that any one of these propositions has not been proved beyond a reasonable doubt, you should find the defendant not guilty. (Illinois Pattern Jury Instructions, Criminal, § 14.05 and 14.06, 1981)

The defining features specified by the Illinois instructions also appear in the Model Penal Code definition of burglary, which has been used as a guide by many states when drafting their criminal codes (American Law Institute, 1962; see also Black,
1983). These preliminary plus burglary subjects had access to both their prior knowledge of burglary and the judge’s instructions. This condition replicated previous research on the effectiveness of jury instructions and was expected to produce a similar result, namely that subjects who hear the substantive instructions on burglary make the same decisions as those who hear no legal definition, voting guilty more often for typical than atypical category members (Smith, 1991b).

In the remaining two instruction conditions, subjects heard both the preliminary instructions and the substantive definition of burglary, but the category label burglary was omitted. Subjects were told that the defendants were charged with a serious crime and they were instructed on the definition of that crime without reference to the term burglary. The first sentences of the substantive definition were changed to “A person commits the crime charged when he, without authority, knowingly enters a building with the intent to commit a felony therein. To sustain the charge, the State must prove the following propositions. . . .” The goal of this preliminary plus crime intervention was to prevent subjects from accessing their prior knowledge of burglary by withholding the crime name. To determine whether these subjects access their prior knowledge anyway, they were asked to specify the name of the crime they thought the judge had described. Half the subjects were asked to provide this category label before making any verdict decisions (preliminary plus crime—label before subjects) and half provided a label after making all their verdict decisions (preliminary plus crime—label after subjects). If subjects do not search for a category label until they are asked to do so, preliminary plus crime—label after subjects should be unaffected by the typicality of the scenarios. The information with which they determine the typicality of a fact situation has not been accessed, so they should show no greater tendency to convict on typical than atypical scenarios. Furthermore, if they understand and use the judge’s instructions, the conviction rate should be high for all category members. In contrast, preliminary plus crime—label before subjects are forced to activate their own prior knowledge of a category, and their subsequent decisions will likely be influenced by their category choice. It is possible, of course, that subjects in both preliminary plus crime conditions will search for a category label after hearing the substantive definition of the crime. If so, then both groups will label the crime before making their decisions, and their labels and verdict choices should be similar.

Scenarios

This experiment assessed subjects’ abilities to make two types of categorization decisions: identify true category members and reject non-category members. Six scenarios were presented to test people’s abilities to identify true category members. Each scenario briefly described a fact situation in which the legal requirements for burglary were met. Four of the scenarios contained many of the characteristic features of burglary that subjects in previous research had identified. These scenarios were rated by 100 student judges as quite typical of the category burglary, averaging 6.0 on a 7-point scale where 7 = very typical. For
example, in the following scenario, Hal is guilty of burglary because he entered a building unlawfully with the intent to commit a felony. The characteristic features of the scenario are underlined:

Hal got a tip from a friend that the Andersons were going on vacation for a week, starting December 19th. On the night of the 20th, Hal picked the lock on the Andersons’ back door, and went inside. He dismantled the stereo equipment, the video recorder, and the tv, and carried them out to his car. He looked around to make sure no one had seen him, then went back inside. He went upstairs to the bedroom and filled his pockets with jewelry. He left the house the way he came in, got in his car and drove away.

Two other burglary scenarios contained relatively few characteristic features of burglary and were perceived by 100 student judges to be relatively atypical of burglary, averaging 2.4 on a 7-point scale where 1 = not at all typical. For example, in the following scenario, Art is guilty of burglary because he entered a building unlawfully with the intent to commit a felony, in this case arson. The characteristic features of the scenario are underlined:

Art was fired from his job as a deliveryman, and vowed to get even with his former employer. Early one Saturday morning, Art took a can of gasoline and a pile of rags to the warehouse where he had worked. The warehouse was closed, but Art knew of a window in the back that didn’t lock. He pushed the window open and climbed inside. He spread the gasoline around the warehouse and lit the rags on fire. Then he went home and waited to hear about the fire on the news. However, the fire burned itself out without damaging any property.

In total, there were six scenarios that met the legal requirements for burglary. An experienced trial attorney reviewed these scenarios and confirmed that all six defendants are guilty of burglary under the law. The scenarios varied in how typical of burglary students perceived them to be, with four typical and two atypical crimes. If subjects’ prior knowledge influences their verdict decisions, they should vote guilty more often for the typical than the atypical burglary scenarios. If subjects understand and use the law, they should vote guilty for all six scenarios because all are category members.

To test people’s abilities to reject non-category members, several scenarios were presented that did not meet the legal requirements for burglary; these scenarios described thefts and robberies. For example, in the following scenario Roger is not guilty of burglary because there was no unlawful entry of a building:

Roger was in a park one Saturday afternoon when he noticed a portable cassette recorder propped up against a tree. Roger looked around, but there was no one in sight. He sat down on a bench and waited for several minutes, but no one appeared. Finally, he walked over and picked up the cassette recorder. He looked around again to make sure no one was watching, then walked home with it as fast as he could.

In total, there were six scenarios that did not meet the legal requirements for burglary, two described robberies and four described thefts. An experienced trial attorney reviewed these scenarios and confirmed that none of the six defendants are guilty of burglary under the law. If subjects understand and use the law, these scenarios should produce not guilty votes. The nonburglary scenarios extend previous research on the role of jurors’ prior knowledge in verdict selection; that research did not systematically evaluate subjects’ abilities to reject non-category
HELPING JURORS USE THE LAW

members (Smith, 1991b). Thus, including non-category members in this and the following experiments provides a more complete picture of jurors’ decision-making abilities.

The scenarios were presented to subjects in one of three random orders. After reading each scenario, subjects chose a verdict. For preliminary only and preliminary plus burglary subjects, the verdict question was: “Is (defendant) guilty of burglary?” For subjects in the two preliminary plus crime conditions, the verdict question was: “Is (defendant) guilty of the crime the judge described?” The name of the particular defendant was inserted for each scenario, and subjects responded by checking yes or no. Preliminary plus crime subjects were also asked to provide a category label for the crime: “What do you think is the name of the crime the judge described?” The preliminary plus crime—label before subjects answered this question before they read any of the scenarios or made any decisions; the preliminary plus crime—label after subjects answered this question after reading and choosing verdicts for all 12 scenarios.

Results and Discussion

The goal of this experiment was to circumvent the conflict between subjects’ prior knowledge of burglary and the judge’s instructions by withholding the category label. If successful, preliminary plus crime—label after subjects should not access their prior knowledge of burglary and should rely on the judge’s instructions for verdict selection. These subjects should then be quite accurate at identifying both the typical and atypical burglaries and at rejecting the nonburglaries.

Subjects’ verdict choices on the true burglary scenarios (i.e., those that met the legal requirements for burglary) were submitted to an Instruction × Typicality (typical, atypical) ANOVA, with Typicality as a within-subject factor. This analysis revealed significant main effects of Typicality, $F(1,81) = 133.7, p < .0001$, and Instruction, $F(3,81) = 6.75, p < .001$, and a significant interaction, $F(3,81) = 6.87, p < .001$. As shown in the top two rows of Table 1, subjects in all four instruction conditions performed equally well on the typical burglaries, $F(3,81) = 0.57, n.s.$, but there were differences across instruction conditions on the atypical burglaries, $F(3,81) = 6.87, p < .001$. Newman–Keuls post hoc comparisons revealed that the performance of preliminary plus burglary subjects did not differ significantly from that of preliminary only subjects. Consistent with previous research, hearing the judge’s instructions on burglary did not significantly improve subjects’ abilities to identify atypical category members. However, subjects in both preliminary plus crime conditions did outperform the preliminary only subjects, $p$’s < .05; hearing the substantive instruction with the crime name un-

---

2 The order of presenting scenarios interacted significantly with typicality, $F(2,73) = 5.92, p < .01$. For all three orders, typical scenarios produced significantly more guilty votes than atypical scenarios, as expected, all $p$’s < .0001, but the difference was larger for Order 2 (1.00 vs. .31) than for Orders 1 (1.00 vs. .62) and 3 (.98 vs. .50). Because the typicality effect was in the same direction and significant for all three orders, the data were collapsed across order. Neither the Order × Instruction nor the Order × Typicality × Instruction interaction was significant. Unless otherwise indicated, there were no order effects in the remaining analyses.
Table 1. Conviction Rates on Burglary and Nonburglary Scenarios in Each Instruction Condition

<table>
<thead>
<tr>
<th>Scenario type</th>
<th>Preliminary only</th>
<th>Preliminary plus burglary</th>
<th>Preliminary plus crime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Burglaries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typical</td>
<td>0.99</td>
<td>1.00</td>
<td>0.99</td>
</tr>
<tr>
<td>Atypical</td>
<td>0.22a</td>
<td>0.39a,b</td>
<td>0.69b, 0.65b</td>
</tr>
<tr>
<td>Nonburglaries</td>
<td>0.77a</td>
<td>0.11b</td>
<td>0.10b, 0.15b</td>
</tr>
</tbody>
</table>

Note. Numbers with different subscripts in a given row differ at p < .05. Rows without subscripts contain no significant differences.

a High numbers reflect greater accuracy. b Low numbers reflect greater accuracy.

specified did improve subjects' abilities to recognize the atypical category members.

The improved performance of preliminary plus crime subjects is consistent with the goal of this experiment, namely, to circumvent the conflict between people's prior knowledge and the law by not activating their prior knowledge and thus forcing them to rely on the judge's instructions for decision making. Nevertheless, there are reasons to believe that this preliminary plus crime intervention did not operate as planned. First, preliminary plus crime subjects did not significantly outperform preliminary plus burglary subjects, although the differences were marginally significant, pairwise .05 < p's < .10. Thus, it is not clear in this experiment that avoiding the category label is a better method of instruction than providing the label, although there is a trend in that direction.

Second, there is evidence that in the preliminary plus crime—label after condition subjects did access their prior knowledge. In this condition, typicality influenced decision making, with typical scenarios producing significantly more guilty votes than atypical scenarios, $F(1,19) = 15.26, p < .001$. If subjects did not access their prior knowledge, there should have been no greater tendency to convict for typical than atypical burglaries; all are category members and all should have resulted in guilty verdicts.

Third, subjects in both preliminary plus crime conditions made similar decisions. The preliminary plus crime—label before subjects were forced to activate their own prior knowledge of a category by labeling the crime before making their verdict decisions. Thus, both the judge's instructions and subjects' prior knowledge of whatever category they chose were potentially available for decision making. The preliminary plus crime—label after subjects were asked to label the crime only after all of their verdict decisions were made. It was hoped that these subjects would refrain from choosing a label for the crime until they were asked to do so, and thus have only the judge's instructions available for decision making. If this occurred, the label before and label after groups would have access to different information for verdict selection and should have made different deci-
sions. However, the verdict choices of these two groups did not differ, suggesting that they were using similar decision processes. This conclusion is further supported by the labels subjects chose in response to the question “What do you think is the name of the crime the judge described?” The distribution of label choices did not differ for the label before and label after conditions, \( \chi^2(2, N = 41) = 0.42, \) n.s. Both groups chose the label *breaking and entering* most frequently (44% for label after and 48% for label before subjects), followed by *burglary* (22% and 14%, respectively). Other labels, including *stealing, theft*, and *robbery*, were the remaining choices (34% and 38%, respectively).

The similarity in label choices and verdict decisions for both preliminary plus crime conditions suggests that the two groups were using similar decision processes. It appears that after hearing the judge’s instructions, subjects selected a label for the unspecified crime, whether they were asked to do so or not. This gave them access to their prior knowledge of the category selected, which they could then use for decision making. The fact that both preliminary plus crime conditions voted guilty more often for the typical than the atypical burglaries, \( p’s < .002, \) indicates that this prior knowledge was influential. Under these circumstances, there are two possible reasons that the preliminary plus crime subjects were more accurate than the preliminary only subjects when identifying atypical category members. First, preliminary plus crime subjects may have been unwilling to rely on their own guesses about what crime was charged. Uncertainty about the correctness of their chosen label may have prompted them to rely less on their prior knowledge and more on the judge’s instructions, thus increasing their accuracy. Preliminary only and preliminary plus burglary subjects were told that the defendants were charged with burglary, so there was no uncertainty for them about what crime was charged and they could be confident about what prior knowledge to access. If this characterization of subjects’ decision making is correct, then withholding the crime name may be a useful intervention. It may not avoid the conflict between people’s prior knowledge and the law, but it may discourage their reliance on their prior knowledge. If so, then preliminary plus crime subjects should outperform preliminary only subjects regardless of the target crime. Experiment 2 tested this possibility using kidnapping cases.

The second possible reason that preliminary plus crime subjects outperformed preliminary only subjects on the atypical burglaries is that the former made a fortunate choice of category labels. About half of the preliminary plus crime subjects labeled the crime *breaking and entering*. As all of the burglary scenarios involved unlawful entry, using this label as a basis for verdict selection could indeed boost accuracy on the atypical burglaries. Thus, the improvement in performance of the preliminary plus crime subjects may not reflect an understanding and use of the judge’s instructions; it may be a product of subjects’ label choices. If this is true, preliminary plus crime subjects who chose the label *breaking and entering* should be significantly more accurate at identifying atypical category members than subjects who chose other labels.

To test this possibility, the verdict choices of the preliminary plus crime subjects were submitted to a 3-way ANOVA with Instruction condition (label before versus label after) and Label (breaking and entering versus other) as be-
tween-subjects factors and Typicality as a within-subject factor. The only significant effect was a Label × Typicality interaction, $F(1,35) = 4.35, p < .05$. As predicted, subjects who labeled the crime breaking and entering were more accurate (typical = .99, atypical = .82) than subjects who chose other labels (typical = 1.00, atypical = .60) when it came to identifying atypical category members.\(^3\) These results suggest that the improved performance of preliminary plus crime subjects may have been a byproduct of a fortunate choice of labels. If this explanation is correct, such improvement is unlikely to replicate with a different target crime. Experiment 2 tested this prediction using kidnapping cases.

There was a significant effect of typicality in each of the four instruction conditions, all $p$'s < .005. This suggests that, as predicted, subjects' prior knowledge did influence their decision making, despite important variations in the instructions they heard. It is possible, however, that the differences in conviction rates for the typical and atypical scenarios were not due to variations in typicality, but to differences in (a) the perceived heinousness of the defendants' actions or (b) the degree to which the legal elements of burglary were met. The first of these possible alternative explanations suggests that subjects voted guilty more often for the typical than the atypical scenarios because the former described more heinous crimes. To test this possibility, 20 new subjects read the typical and atypical burglaries and rated for each one how heinous the defendant's actions were on a 7-point scale, where 1 = not at all heinous and 7 = extremely heinous. A matched-pair t test revealed that the atypical scenarios were perceived to be just as heinous as the typical scenarios (4.58 vs. 4.70, respectively), $t < 1$, n.s., indicating that the difference in conviction rates for typical and atypical crimes was not a byproduct of perceived heinousness.

The second alternative explanation suggests that subjects voted guilty more often for the typical than the atypical crimes because they believed that the legal requirements for burglary were met in the typical scenarios but were not met in the atypical scenarios. To test this possibility, 20 additional subjects judged for each scenario whether the legal requirements were met. Subjects were asked (1) did (defendant) knowingly enter a building?, and (2) if yes, did he enter without authority?, and (3) did he enter with the intent to commit a felony? The name of the particular defendant was inserted for each scenario, and subjects responded to each question by checking yes or no. If subjects answered yes to all three questions, then they believed that all of the legal requirements for burglary were met; a no response on any one of these questions meant they believed the legal requirements for burglary were not met. Subjects answered yes to all three questions equally often for the typical and atypical burglaries (.93 vs. .88, respectively), $F < 1$, n.s., indicating that both types of scenarios satisfied the legal requirements for burglary. This, of course, means that subjects should have voted

\(^3\) A similar analysis was conducted comparing subjects who chose the label breaking and entering and those who chose the label burglary. Although the means were in the right direction (atypical breaking and entering = .82, atypical burglary = .57), the Label × Typicality interaction did not reach significance, $p = .12$, probably because of insufficient power; only 7 subjects chose the label burglary.
guilty at a high rate for both types of burglaries. However, the verdict choices in Experiment 1 differed dramatically from this pattern—subjects voted guilty at a high rate for the typical burglaries, but not for the atypical burglaries.

Also of interest in the present experiment was the impact of the judge’s instructions on subjects’ abilities to reject non-category members, an issue that previous research did not systematically address (Smith, 1991b). Verdict choices on the nonburglary (robbery and theft) scenarios were submitted to a one-way ANOVA, which yielded a significant effect of Instruction, $F(3,81) = 35.6, p < .0001$. As illustrated in the bottom row of Table 1, the conviction rate was higher for subjects who heard only preliminary instructions than for subjects who heard the substantive definition of the crime charged. This higher conviction rate means that preliminary only subjects made more errors than subjects in the other conditions, pairwise $p's < .05$ by Newman–Keuls tests, which did not differ significantly from each other. This same pattern of results was obtained when the two types of nonburglaries (theft and robbery) were analyzed separately. Thus, hearing the definition of the crime charged helped subjects to reject non-category members, regardless of whether the crime name was specified. In fact, preliminary only subjects were significantly less likely to vote guilty on the atypical burglaries than on the nonburglaries, $F(1,24) = 42.2, p < .0001$. Hearing the definition of burglary favorably changed that pattern. Preliminary plus burglary subjects were marginally more likely to vote guilty on the atypical burglaries than on the nonburglaries, $F(1,18) = 4.04, p = .06$, and preliminary plus crime subjects were significantly more likely to do so, both $p's < .001$.

The improvement on the nonburglaries among preliminary plus crime subjects may be due to their tendency to label the crime breaking and entering. Although some of the thefts and robberies were committed indoors (e.g., in a department store, locker room, bar), none involved a break-in. Correctly rejecting these scenarios, then, may have been a product of subjects’ label choice. Indeed, subjects who chose the label breaking and entering were significantly more accurate at rejecting the non-category members than were subjects who chose other labels (.04 vs. .20), $F(1,35) = 4.58, p < .05$. However, this explanation cannot account for the unexpected improvement of the preliminary plus burglary subjects. Hearing the substantive instructions on burglary did not help these subjects to correctly identify atypical burglaries, but it did help them to correctly reject nonburglaries. The reason for this selective improvement is unclear. It may be that subjects extracted a partial understanding of the category definition from the instructions. If, for example, subjects learned that unlawful entry is important, but failed to learn the intent to commit a felony feature, they could reject many nonburglaries more accurately without being able to identify atypical burglaries. Taken together, the results for the preliminary plus burglary subjects indicate that the instructions were not wholly effective in educating subjects about burglary, but it appears that they were partially effective. Experiment 2 also investigated subjects’ abilities to reject non-category members to determine whether this selective improvement replicates with a different target crime.

Differences in the patterns of results for the burglaries and nonburglaries were not due to differences in perceived heinousness of these crimes; independent
ratings indicated that subjects perceived the nonburglaries to be just as heinous as the burglaries (4.9 vs. 4.65), t < 1, n.s. Furthermore, 20 additional subjects verified that the legal requirements for burglary were not met in the nonburglary scenarios. Subjects were much less likely to believe that all three legal elements were satisfied for the nonburglaries (.03) than for either the typical (.93) or the atypical (.88) true burglaries, both p's < .001. Thus, subjects believed that the nonburglaries did not satisfy the legal requirements for burglary, and they should have voted not guilty.

EXPERIMENT 2

The purpose of Experiment 2 was to determine whether the improvements in decision accuracy obtained in Experiment 1 replicate with a different target crime, namely, kidnapping. If the preliminary plus crime subjects in Experiment 1 were better able to identify atypical category members than the preliminary only subjects because of either genuine gains in their understanding and use of the law or a decreased reliance on their prior knowledge, then the same improvement should be obtained with a new crime. However, if their improved performance was due to a fortuitous choice of labels for the target category, then the improvement is unlikely to replicate. This experiment also explores more comprehensively subjects' abilities to reject non-category members. Both typical and atypical non-category members were included to provide more information about the role of people's prior knowledge in these decisions.

Method

Subjects and Procedure

Subjects were 80 introductory psychology students who participated in partial fulfillment of a course requirement. Subjects were randomly assigned to one of the four instruction conditions of Experiment 1. They listened to an audiotape of the judge's instructions, then made verdict decisions for eight scenarios. At the conclusion of the experimental session, subjects were fully debriefed.

Instructions

Preliminary only subjects were told that the defendants were charged with kidnapping, then they heard the preliminary instructions on the presumption of innocence and reasonable doubt. Preliminary plus kidnapping subjects were told the defendants were charged with kidnapping, then they heard the preliminary instructions and the substantive definition of kidnapping:

A person commits the offense of kidnapping when he knowingly and (1) secretly confines another person against his will; or (2) by force, or by threat of imminent force, carries another person from one place to another with intent secretly to confine that person against his will; or (3) by deceit or enticement induces another person to go from one place to another place with intent secretly to confine that person against his will. To
sustain the charge of kidnapping, the State must prove the following propositions: first, that the defendant acted knowingly; and second, that the defendant secretly confined the victim against his will; or second, that the defendant, by force or threat of imminent force, carried the victim from one place to another place; and third, that when the defendant did so he intended secretly to confine the victim against his will; or second, that the defendant, by deceit or enticement, induced the victim to go from one place to another place; and third, that when the defendant did so he intended secretly to confine the victim against his will. If you find from your consideration of all the evidence that each one of these propositions has been proved beyond a reasonable doubt, you should find the defendant guilty. If you find from your consideration of all the evidence that any one of these propositions has not been proved beyond a reasonable doubt, you should find the defendant not guilty. (Illinois Pattern Jury Instructions, Criminal, § 8.01 and 8.02, 1981)

The defining features of kidnapping specified in the Illinois instructions also appear in the Model Penal Code definition, which has been used by many states when drafting their criminal codes (American Law Institute, 1962; see also Black, 1983). Preliminary plus crime subjects were told that the defendants were charged with a serious crime, then they heard the preliminary instructions and the substantive definition of kidnapping, with the word kidnapping removed.

Scenarios

As in Experiment 1, subjects' abilities to perform two kinds of categorization decisions were of interest: identify true category members and reject non-category members. The true category members met the legal requirements for kidnapping, but varied in how typical of kidnapping subjects perceived them to be. Two scenarios contained many characteristic features and were rated by 100 students as quite typical, averaging 5.75 on a 7-point scale where 7 = very typical. For example, in the following scenario Ken is guilty of kidnapping under the law because he secretly confined Tony against his will. The characteristic features are underlined:

Tony was playing ball with his friends in the playground of his elementary school one afternoon. Ken knew that Tony’s parents were very wealthy and very protective of their child. Ken called to Tony and waved him over to his car. When Tony came over, Ken asked if he wanted to go for ice cream. Tony said, “Sure,” and got in the car. Ken bought Tony an ice cream, then took him to a motel room, where he tied him to a chair, gagged him, and told him if he made any noise he’d kill him. Ken then took a note demanding $500,000 for Tony’s safe return, and left it in the mailbox of Tony’s parents’ house.

Two of the true kidnapping scenarios contained few characteristic features and were rated by 100 students as relatively atypical, averaging 2.87 where 1 = not at all typical. For example, in the following scenario the drug dealers are guilty of kidnapping because they secretly confined Leon against his will:

Leon was an investigator with the Drug Enforcement Administration. He discovered the headquarters of a major drug ring and went there to gather evidence against them. He hid behind a stack of boxes and listened to their conversation. In their discussion, the dealers revealed everything Leon needed to bust them. When he tried to sneak out, he knocked over one of the boxes and was discovered. The dealers told Leon to tell them everything
he knew or they would beat him until he did. Leon wouldn’t tell them anything, so the dealers locked him in their hide-out and left him there.

An experienced trial attorney reviewed these scenarios and confirmed that the defendants in all four of the true kidnapping scenarios are guilty of kidnapping under the law.

There were four scenarios that did not meet the legal requirements for kidnapping. These non-category members also varied in how typical of kidnapping subjects perceived them to be. Two of the non kidnappings contained many characteristic features and were rated by 64 students as moderately typical of kidnapping, averaging 3.92 where 7 = very typical. For example, in the following scenario, Neil is not guilty of kidnapping because he did not secretly confine the child or intend to secretly confine her.

Neil and his wife went through a nasty divorce and custody battle. Neil’s wife was awarded custody of their daughter, but she got to stay with Neil every third weekend of the month. One Friday afternoon, Neil drove to his daughter’s elementary school. She was playing with her friends in the schoolyard. Neil waved to her and she got in the car with him. Neil took her to his apartment briefly, and then the two went to a basketball game. It was only the second weekend of the month, and Neil’s wife did not know that Neil had picked up the child from school. The girl’s friends said only that she had gotten into a car with some man. Neil’s wife reported the incident to the police.

Two of the non kidnappings contained few characteristic features and were perceived by 64 students as relatively atypical of kidnapping, averaging 2.55 where 1 = not at all typical. For example, in the following scenario Craig is not guilty of kidnapping because Jennifer consented to go away with him:

Jennifer was planning to go away for the weekend with her secret lover, Craig. She agreed to meet Craig downtown, and they would go to Jamaica together. Craig drove up to the designated place and got out of the car. Jennifer was walking down the street when Craig walked over to her, took her arm, and escorted her quickly to the car. Jennifer looked around to see if anyone was watching, then got into the car. One of her husband’s employees saw this incident, and became alarmed. He wrote down the license plate of the car, and went back to work to tell his boss what had happened. Jennifer’s husband called the police and told them his wife was missing. Meanwhile, Jennifer and Craig drove to the airport and left the country together.

The typical non kidnappings were rated as only moderately typical (M = 3.92), but they were perceived as significantly more typical than the atypical non kidnappings (M = 2.55), t(63) = 7.53, p < .0001. An experienced trial attorney verified that none of the four defendants in the non kidnapping scenarios is guilty of kidnapping under the law. Thus, if subjects understand and use the judge’s instructions, all of these scenarios should produce not guilty verdicts because none meets the legal requirements for kidnapping. In contrast, if subjects’ prior knowledge of kidnapping influences their decisions, typical non kidnappings should produce more guilty votes than atypical non kidnappings.

The scenarios were presented to subjects in one of three random orders. For each scenario, preliminary only and preliminary plus kidnapping subjects were asked “Is (defendant) guilty of kidnapping?” Preliminary plus crime subjects were asked “Is (defendant) guilty of the crime the judge described?” The name of
the particular defendant was inserted for each scenario and subjects responded by checking yes or no. Preliminary plus crime subjects were also asked "What do you think is the name of the crime the judge described?" As in Experiment 1, half the subjects answered this question before making any of their verdict decisions, half answered it after making all of their verdict decisions.

Results and Discussion

The goal of this experiment was to determine whether withholding the name of the crime category effectively prevents subjects from accessing their prior knowledge of kidnapping and thus enhances their abilities to identify atypical category members. Subjects’ verdict choices on the true kidnappings were submitted to an Instruction × Typicality ANOVA, with Typicality as a within-subject factor. This analysis revealed a significant main effect of Typicality, in which typical scenarios produced a higher conviction rate than atypical scenarios (.96 vs. .69, respectively), $F(1,76) = 41.3, p < .0001$. Thus, over all four instruction conditions, subjects’ prior knowledge did influence their verdict decisions. There was no main effect of Instruction, $F(3,76) = 0.85, n.s.$, and no interaction, $F(3,76) = 0.80, n.s.$ As illustrated in the top two rows of Table 2, subjects in all conditions performed equally well on the typical kidnappings and equally poorly on the atypical kidnappings.

Consistent with Experiment 1, preliminary plus kidnapping subjects made the same decisions as preliminary only subjects. Unlike Experiment 1, preliminary plus crime subjects did not outperform the preliminary only group, pairwise $p$’s > .05 by Newman-Keuls tests; neither did they outperform the preliminary plus kidnapping group, indicating that withholding the crime name did not significantly enhance decision accuracy on the kidnapping scenarios. The preliminary plus crime—label before and preliminary plus crime—label after conditions produced comparable conviction rates, suggesting that these two groups were using similar

<table>
<thead>
<tr>
<th>Instructions</th>
<th>Preliminary only</th>
<th>Preliminary plus kidnapping</th>
<th>Preliminary plus crime label before</th>
<th>Preliminary plus crime label after</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kidnappings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typical</td>
<td>0.95</td>
<td>0.98</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>Atypical</td>
<td>0.60</td>
<td>0.73</td>
<td>0.65</td>
<td>0.78</td>
</tr>
<tr>
<td>Nonkidnappings</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typical</td>
<td>0.45&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.15&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.23&lt;sup&gt;a,b&lt;/sup&gt;</td>
<td>0.28&lt;sup&gt;a,b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Atypical</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note. Numbers with different subscripts in a given row differ at $p < .05$. Rows without subscripts contain no significant differences.

<sup>a</sup> High numbers reflect greater accuracy.

<sup>b</sup> Low numbers reflect greater accuracy.
decision processes. Subjects in both conditions overwhelmingly labeled the crime *kidnapping* (90% of label before subjects and 95% of label after subjects), $\chi^2(1, N = 38) = 1.03$, n.s. Both groups’ verdicts showed a significant effect of Typicality, $F(1,19) = 12.7$, $p < .01$ for label before subjects and $F(1,19) = 4.41$, $p < .05$ for label after subjects, indicating that their prior knowledge did influence their decisions. This similarity in labels and verdict choices suggests that both groups actually labeled the crime before making their decisions and accessed their prior knowledge of kidnapping for verdict selection. The fact that their decisions did not differ from the preliminary only and preliminary plus kidnapping groups suggests that they relied on this prior knowledge just as heavily as subjects who were told explicitly what crime was charged. Together, these findings reveal no improvement in performance when the crime name is withheld. Thus, it appears that the gains in identifying atypical category members obtained in the previous experiment were not the result of a successful intervention.

A significant typicality effect emerged in each of the four instruction conditions, $p$'s < .05, indicating that subjects’ prior knowledge did influence decision making, even with important variations in instructions. It is possible that the differences in conviction rates for the typical and atypical scenarios are due to differences in perceived heinousness or the degree to which the legal elements of kidnapping were met. To test the heinousness hypothesis, 20 new subjects read the typical and atypical kidnappings and rated for each one how heinous the defendant’s actions were on a 7-point scale (1 = *not at all* heinous and 7 = *extremely* heinous). A matched-pairs $t$ test revealed that the atypical scenarios were perceived to be just as heinous as the typical scenarios (5.88 vs. 6.15), $t < 1$, n.s., indicating that the difference in conviction rates is not a byproduct of perceived heinousness.

To test the possibility that subjects believed the legal elements for kidnapping were met for the typical but not met for the atypical crimes, 20 new subjects judged for each scenario whether the legal elements of kidnapping were satisfied. They were asked (1) did (defendant) secretly confine (victim)? If yes, did he confine (victim) against his/her will?; (2) did (defendant) take (victim) from one place to another by force or threat of force? If yes, did he intend to secretly confine (victim)? Did he intend to confine (victim) against his/her will?; (3) did (defendant) induce (victim) to go from one place to another by deceit or enticement? If yes, did he intend to secretly confine (victim)? Did he intend to confine (victim) against his/her will? The names of the particular defendant and the particular victim were inserted for each scenario. Subjects responded to each question by checking yes or no. Subjects indicated that the legal requirements for kidnapping were met by (a) answering yes to both parts of Question 1, and/or (b) answering yes to all three parts of Question 2, and/or (c) answering yes to all three parts of Question 3. Any other pattern of responses indicated that the subject did *not* believe the legal elements of kidnapping were satisfied. In this experiment, subjects indicated that the legal requirements were met by both the typical and the atypical kidnappings (.95 vs. 1.00), $F = 1$, n.s. This means that the differences in conviction rates for the typical and atypical scenarios are not due to differing perceptions of the degree to which the legal elements of kidnapping were satisfied.

Subjects’ verdict choices for the noncategory members were also submitted
to an Instruction × Typicality ANOVA. This analysis yielded main effects of Typicality, $F(1,76) = 51.2, p < .0001$, and Instruction, $F(3,76) = 2.75, p < .05$, and a significant interaction, $F(3,76) = 2.75, p < .05$. As illustrated in the bottom two rows of Table 2, subjects in all four conditions performed equally well on the atypical nonkidnappings, but there were condition differences on the typical nonkidnappings, $F(3,76) = 2.75, p < .05$. Preliminary only subjects voted guilty significantly more often than preliminary plus kidnapping subjects, meaning that the preliminary plus kidnapping group was more accurate at rejecting the non-category members, $p < .05$ by Newman–Keuls post hoc tests. Thus, as in Experiment 1, hearing the definition of the target crime improved subjects’ abilities to reject non-category members, suggesting that the instructions were partially effective. The conviction rates of the preliminary plus crime conditions fell between the other groups and did not differ significantly from either.

In each of the four instruction conditions, typical nonkidnappings produced a higher conviction rate than atypical nonkidnappings, $p$'s < .05, indicating that subjects' prior knowledge did influence their decisions about non-category members. If their prior knowledge of kidnapping had played no role in these decisions, the conviction rates for the typical and atypical scenarios would have been equal and ideally would have been equally low. Again, it is possible that the differences in conviction rates for the typical and atypical nonkidnappings are due to differences in the perceived heinousness of the scenarios or the degree to which the legal elements of kidnapping are satisfied. However, 20 subjects rated the heinousness of each nonkidnapping and perceived the atypical scenarios to be just as heinous as the typical scenarios (2.75 vs. 3.28), $t(19) = 1.69$, n.s. Twenty additional subjects were asked to determine whether the legal elements of kidnapping were met. They indicated that both the typical and atypical nonkidnappings failed to meet the legal requirements (.05 vs. .00 satisfied the legal elements), $F(1,19) = 2.11$, n.s. Thus, neither possible alternative explanation can account for the higher conviction rate on the typical nonkidnappings.

It is important to remember that the terms typical and atypical refer to how typical of kidnapping a scenario is perceived to be. The designations true kidnapping and nonkidnapping are independent of typicality and are based on whether or not the scenario meets the legal requirements for kidnapping. Thus, a typical nonkidnapping is a scenario that does not constitute kidnapping under the law, but that subjects nevertheless perceive to be relatively typical of kidnapping.

The partial effectiveness of the judge’s instructions is also demonstrated by a comparison of the middle two rows of Table 2. The atypical kidnappings and the typical nonkidnappings are the two types of scenarios for which subjects’ prior knowledge and the law are most inconsistent. An Instruction × Scenario type (atypical kidnappings vs. typical nonkidnappings) ANOVA revealed no main effect of instruction, $F(3,76) = 0.90$, n.s. However, there was a significant main effect of scenario type, with atypical kidnappings producing more guilty verdicts than typical nonkidnappings, $F(1,76) = 50.9, p < .0001$, and a marginally significant interaction, $F(3,76) = 2.57, p = .06$. Follow-up analyses revealed that, in the preliminary only condition, the conviction rates for the atypical kidnappings and the typical nonkidnappings did not differ (.60 vs. .45), $F(1,19) = 1.31$, n.s. In the other three conditions, in which subjects heard the substantive definition of kidnapping, the atypical kidnappings produced significantly more guilty votes than the typical nonkidnappings, as they should, all $p$'s < .001. As noted earlier, this improvement among instructed subjects is due to more accurate rejections of the typical nonkidnappings.

It is interesting to note that typicality and substantive instructions are not the only factors influencing subjects’ verdict choices in this experiment. As illustrated in Table 2, subjects voted guilty much
Together, the results of Experiments 1 and 2 suggest that we cannot circumvent the conflict between people’s prior knowledge and the law by withholding the name of the crime charged. It appears that people label the crime for themselves and activate their own prior knowledge. In both experiments, subjects in the two preliminary plus crime conditions made comparable verdict decisions and label choices and both exhibited typicality effects, suggesting that both groups labeled the crime and activated their prior knowledge before making their decisions, whether they were specifically asked to do so or not. Thus, it appears that withholding the name of the crime charged is not a viable solution to the conflict between prior knowledge and the law. If this conflict cannot be avoided, perhaps its impact can be minimized by reducing people’s reliance on their prior knowledge for decision making. Current jury instructions inform jurors of the legally correct decision criteria and strategy, but they do not acknowledge jurors’ naive concepts or warn them of possible discrepancies between their prior knowledge and the law. Perhaps specifically informing jurors that their prior knowledge is irrelevant would prompt them to focus their attention more narrowly on the judge’s instructions.

EXPERIMENT 3

Experiment 3 tested the effectiveness of a supplementary instruction designed to discourage subjects’ reliance on their prior knowledge of kidnapping. Subjects were told that they should not base their verdict decisions on their own knowledge of what constitutes kidnapping; they should disregard how typical or atypical of kidnapping the facts seem to be and rely solely on the information

more often for atypical kidnappings than atypical nonkidnappings. Because this occurred in all four instruction conditions, this effect cannot be attributed to delivery of the substantive definition of kidnapping. Furthermore, pretest subjects perceived the atypical kidnappings and the atypical nonkidnappings to be similarly atypical (2.87 vs. 2.55 on a 7-point scale), t(162) = 1.62, n.s. Therefore, the differences in conviction rates cannot be due to differences in perceived typicality either. There are two possible explanations for these conviction differences. First, all subjects heard preliminary instructions on the procedural law and requirements of proof, and they may have extracted some information from those instructions that prompted different verdicts on these scenarios. However, it is not at all clear how general instructions on the presumption of innocence and reasonable doubt (that contain no information about the crime of kidnapping) could produce more guilty verdicts on atypical kidnappings than on atypical nonkidnappings. More plausible is the possibility that there is some aspect of people’s prior knowledge that is not captured by typicality ratings. Perhaps, for example, some features of people’s crime prototypes are weighed more heavily for verdict decisions than for typicality judgments. This could produce the conviction differentials on the atypical kidnappings and atypical nonkidnappings obtained in this experiment (and in Experiments 3 and 4) even though the typicality ratings are similar. This would mean that typicality ratings are not perfect descriptors of people’s prior knowledge and are, therefore, not perfect predictors of verdict choice. This is an important prospect that should be addressed in future research. What is critical for these experiments is that typicality does play an important (although perhaps not perfect) role in subjects’ verdict decisions. Consistent with previous research, these experiments demonstrate that people’s prior knowledge of the law powerfully (and inappropriately) influences their decision making; these experiments strive to reduce that influence.
HELPING JURORS USE THE LAW

contained in the judge’s instructions. This instruction has the potential advantages of (a) alerting subjects that their prior knowledge may be inconsistent with the law and (b) underscoring the importance of the judge’s instructions as the basis for decision making.

Method

Subjects and Procedure

Subjects were 60 introductory psychology students who participated in partial fulfillment of a course requirement. Subjects were randomly assigned to one of three instruction conditions. They listened to an audiotape of the judge’s instructions, then made verdict decisions for the eight kidnapping scenarios used in Experiment 2. Four of the scenarios were true kidnappings, two typical and two atypical. Four scenarios were nonkidnappings under the law, two relatively typical of kidnapping and two relatively atypical. The scenarios were presented to subjects in one of three random orders. At the end of the experimental session, subjects were fully debriefed.

Instructions

The preliminary only and preliminary plus kidnapping conditions were identical to those used in Experiment 2. In a third condition, subjects listened to the preliminary instructions and then heard the following supplementary instruction:

These defendants are charged with the crime of kidnapping. Do not base your verdict decisions on how similar the facts are to your own notion of kidnapping. You must disregard how typical or atypical the facts sound. You must base your verdict decisions on the following instructions.

Subjects in this typicality irrelevant condition then heard the substantive definition of kidnapping. The supplementary instruction was delivered before the substantive definition so that subjects would be alerted to the importance of the legal definition and would have the opportunity to process this information more carefully.

Results and Discussion

Subjects’ verdict choices on the true kidnappings were submitted to an Instruction × Typicality ANOVA. This analysis yielded a significant main effect of Typicality, in which typical scenarios produced a higher conviction rate than atypical scenarios (.94 vs .60), $F(1,57) = 44.3, p < .0001$. No other effects were significant. As illustrated in the top two rows of Table 3, all three instruction groups performed equally well on the typical kidnappings and equally poorly on the atypical kidnappings.

An ANOVA on subjects’ verdict choices for the non-category members yielded significant main effects of Typicality, $F(1,57) = 30.4, p < .0001$, and Instruction, $F(3,76) = 4.59; p < .01$, and a significant interaction, $F(3,76) = 5.8, p < .002$. As illustrated in the bottom two rows of Table 3, subjects in all three
Table 3. Conviction Rates on Kidnapping and Nonkidnapping Scenarios in Each Instruction Condition

<table>
<thead>
<tr>
<th>Scenario type</th>
<th>Instructions</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Preliminary only</td>
<td>Preliminary plus kidnapping</td>
<td>Typicality irrelevant</td>
</tr>
<tr>
<td>Kidnappings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typical</td>
<td>0.98</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>Atypical</td>
<td>0.55</td>
<td>0.60</td>
<td>0.65</td>
</tr>
<tr>
<td>Nonkidnappings</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Typical</td>
<td>0.45&lt;sup&gt;a&lt;/sup&gt;</td>
<td>0.13&lt;sup&gt;b&lt;/sup&gt;</td>
<td>0.10&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Atypical</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

*Note. Numbers with different subscripts in a given row differ at *p* < .05. Rows without subscripts contain no significant differences.

<sup>a</sup> High numbers reflect greater accuracy.

<sup>b</sup> Low numbers reflect greater accuracy.

instruction conditions performed equally well on the atypical nonkidnappings. On the typical nonkidnappings, preliminary only subjects made more errors than subjects who heard the definition of kidnapping, as evidenced by a higher conviction rate on these non-category members, pairwise *p*’s < .05 by Newman–Keuls tests. As in the previous experiments, hearing the substantive definition of the target crime improved subjects’ abilities to reject non-category members, even though it did not improve their abilities to identify true category members. Again, it appears that the judge’s instructions were at least partially effective.<sup>7</sup>

These results indicate that the supplementary instruction did not effectively discourage subjects from using their prior knowledge of kidnapping and focus their attention more narrowly on the judge’s instructions. The verdict decisions of the typicality irrelevant subjects did not differ from those of the preliminary plus kidnapping subjects. It should be noted that preliminary plus kidnapping subjects were quite accurate when rejecting nonkidnappings, so there was little room for improvement with the typicality irrelevant instruction. Whether improvement would have occurred had the error rates in the preliminary plus kidnapping condition been higher is thus still an open question. However, there was ample room for improvement on the true kidnappings and no improvement occurred. Therefore, these data provide no evidence that the typicality irrelevant instruction has any benefit for verdict accuracy.

Although this intervention would have provided a simple and straightforward

---

<sup>7</sup> This conclusion is bolstered by comparison of the middle two rows of Table 3. An Instruction × Scenario type (atypical kidnappings vs. typical nonkidnappings) ANOVA revealed no main effect of instruction, *F*(2,57) = 1.88, n.s., a significant main effect of scenario type, *F*(1,57) = 30.1, *p* < .0001, and a significant interaction, *F*(2,57) = 4.14, *p* < .05. Follow-up analyses revealed that preliminary only subjects voted guilty equally often for the atypical kidnappings and the typical nonkidnappings, *F* < 1, n.s. Subjects in the other two conditions, who heard the substantive definition of kidnapping, voted guilty significantly more often for the atypical kidnappings than for the typical nonkidnappings, as they should, both *p*’s < .001. As explained earlier, this improvement is due to more accurate rejection of the typical nonkidnappings.
solution to the conflict between subjects’ prior knowledge and the law, its success depended on people being extremely flexible decision makers. Subjects were asked to set aside both the content of their naive representations (“Do not base your verdict decisions . . . on your own notion of kidnapping”), and their preferred decision strategy (“You must disregard how typical or atypical the facts sound.”). For instructions to be effective, it may be necessary to provide more specific guidance on how to resolve the inconsistencies between prior knowledge and law. To that end, Experiment 4 tested the effectiveness of a supplementary instruction that provided a specific, feature-by-feature critique of subjects’ naive representations of kidnapping.

EXPERIMENT 4

The goal of Experiment 4 was to revise subjects’ prior knowledge of kidnapping so that the information contained in their representations was legally correct. A supplementary instruction provided subjects with concrete information about which features of their prototypes should be revised, and how. Rather than asking people to set aside a naive representation that they believe to be relevant and useful, this instruction concentrated on revising the misconceptions contained in people’s representations to make them more consistent with the legal definition.

Method

Subjects and Procedure

Subjects were 94 introductory psychology students who participated in partial fulfillment of a course requirement. Subjects were randomly assigned to one of three instruction conditions. They listened to an audiotape of the judge’s instructions, then made verdict decisions on the eight kidnapping scenarios used in Experiment 3. The scenarios were presented in one of three random orders. At the end of the experimental session, subjects were fully debriefed.

Instructions

The preliminary only and preliminary plus kidnapping conditions were identical to those used in Experiments 2 and 3. The third condition introduced a short supplementary instruction that contained a feature-by-feature evaluation of people’s prior knowledge. The features that were included in the supplementary instruction were those that subjects in previous research had most frequently listed as characteristic of kidnapping (Smith, 1991b). Subjects in this preliminary plus features condition were told that the defendants were charged with kidnapping, then they heard the preliminary instructions and the following supplementary instruction:

Many people believe that kidnapping requires a ransom demand. However, a person can be found guilty of kidnapping even when ransom is not demanded, and even when the motive for the crime is not money. It is also not necessary that the victim of the kid-
napping be a child. Adults can be victims of kidnapping as well. When the victim is a child, it is not necessary that the kidnapping result from a custody battle. A person can be found guilty of kidnapping even when that person is not involved in a custody battle. It is also not necessary that the victim of the kidnapping be taken to another location. A person can be found guilty of kidnapping even if the victim is not taken away, as long as the following requirements are met.

After this instruction, subjects heard the substantive definition of kidnapping. The supplementary instruction preceded the legal definition so that subjects would be alerted to the importance of the legal requirements before they were presented.

Results and Discussion

If the preliminary plus features instruction effectively revises subjects’ naive representations of kidnapping, these subjects should be better able to identify atypical category members, producing a conviction rate near 100%. Subjects’ verdict choices on the true kidnappings were submitted to an Instruction × Typicality ANOVA, which revealed main effects of Typicality, F(1,91) = 47.5, p < .0001, and Instruction, F(2,91) = 5.15, p < .01, and a significant interaction, F(2,91) = 8.03, p < .001. As illustrated in the top two rows of Table 4, there were no instruction differences on the typical kidnappings, with all three groups performing equally well, F(2,91) = 1.02, n.s. However, there was a significant effect of instruction for the atypical kidnappings, F(2,91) = 6.94, p < .002. As in the previous experiments, preliminary plus kidnapping subjects performed no better than preliminary only subjects. However, those in the preliminary plus features condition performed significantly better than both of the other groups, pairwise p’s < .05 by Newman–Keuls post hoc tests. In fact, these subjects performed just as accurately on the atypical kidnappings as they did on the typical kidnappings, F(1,30) = 1.35, n.s. In contrast, the preliminary only and preliminary plus kidnapping conditions both exhibited significant typicality effects, both p’s < .001. These results indicate that attacking subjects’ misconceptions on a feature-by-feature basis can markedly improve their ability to identify atypical category

Table 4. Conviction Rates on Kidnapping and Nonkidnapping Scenarios in Each Instruction Condition

<table>
<thead>
<tr>
<th>Scenario type</th>
<th>Instructions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Preliminary only</td>
</tr>
<tr>
<td>Kidnappingsa</td>
<td></td>
</tr>
<tr>
<td>Typical</td>
<td>1.00</td>
</tr>
<tr>
<td>Atypical</td>
<td>0.57a</td>
</tr>
<tr>
<td>Nonkidnappingsb</td>
<td></td>
</tr>
<tr>
<td>Typical</td>
<td>0.39a</td>
</tr>
<tr>
<td>Atypical</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note. Numbers with different subscripts in a given row differ at p < .05. Rows without subscripts contain no significant differences.

a High numbers reflect greater accuracy.

b Low numbers reflect greater accuracy.
members and can do so without impairing their ability to identify typical category members.

Verdict choices for the nonkidnapping scenarios were also submitted to an Instruction × Typicality ANOVA. This analysis revealed significant main effects of Typicality, $F(1,91) = 62.8, p < .0001$, and Instruction, $F(2,91) = 5.86, p < .005$, and a significant interaction, $F(2,91) = 5.86, p < .005$. As illustrated in the bottom two rows of Table 4, subjects in all three instruction conditions performed equally well on the atypical non kidnappings, but there were instruction differences on the typical non kidnappings, $F(2,91) = 5.86, p < .005$. As in the previous experiments, subjects who heard the substantive definition of kidnapping significantly outperformed subjects who heard only preliminary instructions, pairwise $p$’s < .05 by Newman–Keuls post hoc tests, indicating that the substantive instruction was at least partially effective. The preliminary plus features subjects did not perform better than the preliminary plus kidnapping subjects on the typical non kidnappings, indicating that the supplementary instruction did not significantly improve the rate of correct rejections. However, error rates for both groups were quite low, and there was little room for the preliminary plus features intervention to produce additional improvement. Whether such improvement would have occurred had error rates in the preliminary plus kidnapping condition been higher is thus still an open question.

On the nonkidnapping scenarios in this experiment, there was a significant Order of Presentation × Typicality × Instruction interaction, $F(4,85) = 5.02, p < .01$. Follow-up analyses were conducted within each presentation order to pinpoint the source of this interaction. These analyses were not very powerful, with only 10 or 11 subjects in each instruction condition. Nevertheless, there were significant typicality effects for all three orders, $p$’s < .001. The Instruction main effect and Instruction × Typicality interaction did not reach significance for Order 1, although the means were similar to the overall Instruction effect (preliminary only = .23, preliminary plus kidnapping = .11, preliminary plus features = .10), $F(2,28) = 2.12, n.s$. The Instruction main effect and interaction did not approach significance for order 2, with subjects in all three conditions making few errors (preliminary only = .07, preliminary plus kidnapping = .08, preliminary plus features = .14), $F < 1, n.s$. The instruction main effect and interaction were significant for Order 3, with preliminary only subjects making more errors than preliminary plus kidnapping and preliminary plus features subjects (.30, .07, .00), $F(2,28) = 19.27, p < .001$. These analyses suggest that under some circum-

---

Comparison of the middle two rows of Table 4 supports this conclusion. An Instruction × Scenario type (atypical kidnappings vs. typical non kidnappings) ANOVA revealed no main effect of instruction, $F(2,91) = 1.46, n.s.$, a significant main effect of scenario type, $F(1,91) = 103.3, p < .0001$, and a significant interaction, $F(2,91) = 12.0, p < .0001$. In the preliminary only condition, subjects were only marginally more likely to vote guilty for the atypical kidnappings than for the typical kidnappings, $F(1,30) = 3.77, p = .06$. This difference was highly significant for the preliminary plus kidnapping and preliminary plus features subjects, both $p$’s < .0001. As in the previous experiments, subjects who heard the substantive definition were more accurate at rejecting typical non kidnappings. In addition, subjects who heard the supplementary instruction were better able to identify atypical category members.
stances, even subjects who have not heard the substantive instructions can accurately reject non-category members. Overall, however, these subjects tend to make more errors than those who have heard the substantive instructions. It is not clear why order of presentation had an impact in this experiment, but it should be noted that the overall effect of instruction obtained for the non-category members in this experiment is identical to the instruction effects obtained in Experiments 1, 2, and 3, in which order of presentation had no effect on performance. Across experiments, then, this finding of greater accuracy at rejecting non-category members for subjects who heard the substantive instruction is quite consistent.

GENERAL DISCUSSION

These experiments explored potential solutions to the conflict between people’s prior knowledge of crime categories and the legal definitions. Current jury instructions ignore the existence of people’s prior knowledge, and the data have consistently shown that these instructions do not prevent subjects from using their naive concepts for decision making. Preliminary plus burglary and preliminary plus kidnapping subjects in the present experiments were still influenced by the typicality of the fact situations, despite having been instructed about proper legal decision making (see also Smith, 1991b). Apparently, these instructions alone are not sufficient to resolve the conflict between prior knowledge and law.

Experiments 1 and 2 attempted to circumvent the conflict by withholding the name of the crime charged against the defendant. Without this simple retrieval cue, it was hoped that subjects would not be able to access their prior knowledge of the relevant category and would have to rely on the judge’s instructions for guidance when selecting verdicts. The results indicated, however, that subjects who were not told the crime name applied their own category label and accessed their own prior knowledge for decision making. So, not only was this intervention ineffective at resolving the conflict between prior knowledge and law, it has the added danger that jurors may choose the wrong label for the crime and access completely irrelevant prior knowledge. Apparently, then, this intervention does not offer a viable solution.

Experiment 3 investigated the possibility that people could be discouraged from using their prior knowledge of the target crime by informing them explicitly that they must disregard their existing notions of the crime and rely solely on the judge’s instructions for decision making. This is perhaps the most straightforward solution available: If subjects are relying on inappropriate information for verdict selection, inform them of their error so that they might concentrate on learning the correct decision process. However, this supplementary instruction had no effect on decision making; subjects relied just as heavily on their prior knowledge of the target crime when they heard the supplementary instruction as when they did not hear it. This finding is consistent with research on cautionary instructions, in which subjects could not (or did not) follow judges’ admonitions to disregard evidence that was inappropriately presented (Sue, Smith, & Caldwell, 1973; Wolf
& Montgomery, 1977). In this context, instructing subjects to disregard their existing knowledge of a crime category did not prompt them to set aside this seemingly relevant information.

These results suggest that jurors are strongly inclined to use their naive representations of crime categories for decision making, and this prior knowledge cannot simply be avoided or disregarded. Experiment 4 began with the understanding that people draw on their prior knowledge for decision making, and the goal of that experiment was to revise subjects’ existing representations so that they conformed to the requirements of the law. To this end, a supplementary instruction was introduced that attacked subjects’ misconceptions about the target crime on a feature-by-feature basis, giving them specific information about how to revise the features contained in their naive representations. This supplementary instruction produced remarkable improvements in subjects’ abilities to identify atypical category members. In the preliminary plus features condition, subjects’ verdicts for the true kidnappings no longer showed an effect of typicality, with equally high conviction rates for typical and atypical kidnappings. Furthermore, preliminary plus features subjects were just as accurate at rejecting non-category members as preliminary plus kidnapping subjects, suggesting two things. First, this supplementary instruction had no associated costs for decisions that were already made accurately. Second, the improvement in subjects’ abilities to identify atypical category members was not due to a criterion shift, in which subjects simply voted guilty more often for all sorts of fact situations. Preliminary plus features subjects voted guilty more often for atypical category members, but not for non-category members, indicating that the instruction’s effect was selective. It appears, then, that an instruction geared toward revising the misconceptions contained in people’s existing representations is a promising way of improving decision accuracy.

Another interesting feature of the data obtained in these experiments is the selective effectiveness of the substantive instructions defining the crime charged. As in previous research (Smith, 1991b), the present experiments revealed that the substantive instructions did not improve subjects’ abilities to identify true category members. However, instructed subjects were better able to reject non-category members. This improvement was not expected, but was consistently obtained across the four experiments. The reason for the selective effectiveness of the substantive instructions is not clear at this time, but it may reflect the relative ease of making these decisions. Rejecting fact situations may be easier than identifying fact situations as category members because there are multiple decision criteria to consider. Only one of the necessary conditions must be violated to correctly reject a fact situation, but all of the necessary conditions must be met before a crime is correctly identified. So, a decision maker can accurately reject many fact situations with only a partial understanding of the category definition, but cannot succeed in identifying a variety of true category members with only a partial definition. Of course, to correctly reject a wide variety of non-category members, a decision maker must know all of the necessary conditions for the category so that all violations are detected. At that point, the decision maker should know enough about the category requirements to correctly identify the
true category members as well. It is possible that the non-category members used in these experiments were not sufficiently varied to reveal the gaps in subjects' understanding. In sum, the fact that instructed subjects failed to recognize atypical category members indicates that the substantive instructions did not entirely succeed in educating subjects about the law, but the fact that they improved at rejecting non-category members suggests that the substantive instructions were partially effective. Precisely how the substantive instructions modify subjects' representations to produce this selective improvement will be addressed in future research that provides a more sensitive test of how these instructions modify particular features of people's representations.

At this point, it is important to consider the practical implications and limitations of this research. The experiments reported in this article indicate that inconsistencies between people's prior knowledge and the law remain an important obstacle to proper legal decision making even after standard jury instructions are presented. This obstacle cannot simply be avoided or disregarded. Introducing a feature-based supplementary instruction designed to revise people's existing representations appears to be a promising way of improving decision accuracy. However, several important aspects of this intervention remain to be explored in future research. First, it is important to determine the effectiveness of this type of supplementary instruction in more realistic trial situations. Subjects in these experiments made decisions on several scenarios that briefly and unambiguously presented the facts of a criminal encounter. Because this research was concerned with how people integrate their prior knowledge of the law with the judge's instructions when selecting a verdict, it was important that all subjects base their verdict decisions on the same evidence. To that end, the facts were presented in a straightforward, coherent, and internally consistent manner that did not require subjects to interpret the evidence, make credibility judgments, or fill gaps in the testimony. These descriptions differed considerably from the complex, voluminous, and often conflicting evidence that jurors encounter in real trials. This sacrifice of mundane realism was made in the interest of maintaining experimental control and minimizing variability in subjects' verdicts that might result from differing interpretations of the evidence. There is no reason to believe at this time that the existence of evidentiary complexity would alter the way in which jurors integrate their prior knowledge of the law with the judge's instructions. However, this issue remains to be tested in future research. It is possible that the additional demands of evidence processing in a real trial could reduce the salience and impact of a supplementary instruction. Therefore, it is important to assess the power of this feature-based intervention to improve decision accuracy in more complex trial situations.

Second, a more comprehensive picture of the features contained in actual jurors' naive representations of these crimes must be generated. The available research using student subjects offers a clue to what some of those features might be, but to be optimally effective in actual trials, the supplementary instruction must provide a comprehensive critique of the features contained in community members' representations. Third, it is important to investigate the optimal timing of the supplementary instruction. In this research, subjects heard all of the in-
dictions before reading the scenarios and making verdict choices. It is possible that this “pretrial” presentation contributed to the effectiveness of this instruction. Several experiments have demonstrated that jurors actively process and evaluate evidence and testimony throughout a trial (Kassin & Wrightsman, 1979; Smith, 1991a; Weld & Danzig, 1940; Weld & Roff, 1938). Under these circumstances, jurors may select a verdict preference based on their prior knowledge of the crime before they ever hear the judge’s instructions. Presenting the supplementary instruction after trial may have no effect on these preferences because of the difficulty of reevaluating the evidence post hoc. Thus, it is possible that this supplementary instruction will revise jurors’ decision making only when presented pretrial, before the evidence is processed.

Fourth, this feature-based supplementary instruction was written in plain English, and the simplicity of the language may have contributed to its effectiveness (e.g., Charrow & Charrow, 1979; Elwork et al., 1977; Severance & Loftus, 1982). If this kind of supplementary instruction is to be used in actual trials, the resulting language may be more complex than that used in Experiment 4. It is important to determine from judges, attorneys, and legal scholars what kinds of modifications to the instruction would be necessary for its adoption in actual trials and to test the effectiveness of those modified instructions for improving decision accuracy.

The research reported in this article indicates that much can be gained by taking people’s existing concepts of law seriously. Significant improvements in decision accuracy were achieved by providing subjects with information about how to revise their naive representations of crime categories. This does not mean that other reforms of the instruction process are unnecessary. As already discussed, rewriting jury instructions in simpler language and presenting instructions pretrial both have advantages for jurors’ performance. Neither solves the problem of poor comprehension, but each clearly has value. Thus, one might expect greatest improvement in jurors’ performance from a combination of available reforms—pretrial presentation of simply written instructions that both correct the misconceptions about law that jurors bring to trial and explain the proper legal decision criteria.

In summary, the feature-based supplementary instruction of Experiment 4 appears to be a promising method for improving jurors’ use of the law. It acknowledges the existence and influence of people’s naive representations and concentrates on correcting the misconceptions contained in those representations. This emphasis on concept revision, rather than concept formation, appears to be an important step toward effective jury instruction.

REFERENCES


What Do You Expect?: The Influence of People's Prior Knowledge of Crime Categories on Fact-Finding

Vicki L. Smith¹ and Christina A. Studebaker¹,²

Jurors have accumulated prior knowledge of crime categories that influences their verdict choices (V. L. Smith, 1991, 1993). The present research investigates whether this prior knowledge also influences the fact-finding process by operating as an expectancy for information processing. Consistent with the expectancy hypothesis, Study 1 demonstrated that people are more susceptible to misleading information that is perceived to be typical of the crime in question than misleading information that is atypical. Studies 2 and 3 revealed that people's intrusion errors also contain more typical than atypical information. These results indicate that people's prior knowledge of crime categories does indeed serve as an expectancy for fact-finding. Thus, prior knowledge plays a broader role than previously appreciated. The implications of this breadth are discussed.

A juror's duty is to make sense of the facts presented at trial, to listen to the judge's instructions on the law, and then to integrate facts and law into a legally appropriate verdict. As fact-finders, jurors are encouraged to evaluate the evidence and testimony in light of their own observations and life experiences; they are supposed to bring their common sense notions of the world to bear on the fact-finding process. In contrast, jurors' use of the law is highly constrained. They are told that they must follow the law as stated in the judge's instructions.³ Jurors are assumed either to have no prior knowledge of the law or to set that prior knowledge aside when they hear the judge's instructions.

Recent research has revealed that jurors are not blank slates with regard to the law; they have constructed naive representations of crime categories, such as burglary, assault, and kidnapping. In a series of studies, V. L. Smith (1991) found that people were readily able to list features that they believed characterize various

¹Northwestern University. ²Christina Studebaker is now at the Center for Children, Families, and the Law, 121 S. 13th St., Suite 302, P.O. Box 880227, Lincoln, NE 68588-0227. ³Some states instruct jurors that they may nullify the law and vote not guilty if they believe that a guilty verdict would be unjust under the circumstances (Hans & Vidmar, 1986). This instruction is not widely used, and is a rare exception to the expectation that jurors will follow the law as stated by the judge.
crimes, indicating that they had accumulated some prior knowledge of these crimes that they could access when needed. The features they listed tended to be incomplete, incorrect, or irrelevant under the law, indicating that much of their prior knowledge was inconsistent with the law. Several experiments demonstrated that this prior knowledge of crime categories also influences people's verdict decisions (V. L. Smith, 1991, 1993). Participants in these experiments read several short scenarios, each describing the facts of a criminal encounter. Some scenarios contained many characteristic features of the crime (obtained from the earlier participants' feature lists); other scenarios contained few characteristic features. All scenarios met the legal requirements for the crime in question, so all were category members under the law. For each critical scenario, some participants were asked to rate how typical the incident was of the crime in question, others were asked to choose a verdict. As predicted, scenarios containing many characteristic features were rated as more typical of the crime category and produced more guilty verdicts than scenarios containing few characteristic features. Because all of the critical scenarios were category members under the law, all should have produced guilty verdicts. Nevertheless, participants' verdict choices were influenced by their prior knowledge of the crime categories. This influence of prior knowledge persisted even when participants heard the judge's instructions on the law. The verdict choices of instructed and uninstructed participants did not differ; both groups voted guilty more often for typical than atypical category members (V. L. Smith, 1991, 1993). Together, these findings indicate that jurors have accumulated some prior knowledge of crime categories, much of which is erroneous, that influences their verdict choices.

In theory, people's prior knowledge of crime categories might serve several functions. We know from the previous research that prior knowledge influences verdict decisions (V. L. Smith, 1991, 1993). However, it may play a broader role in jurors' decision making, influencing other stages of the decision process as well. The purpose of the present research was to explore the impact of people's prior knowledge of crime categories on the fact-finding process. People bring their prior knowledge to trial with them, so it is potentially available to them throughout the trial. As a result, jurors may use this information as an expectancy for evidence processing. Specifically, their expectations about what is typical for the type of crime charged may influence the inferences they make to fill gaps in the evidence.

Such expectancy effects occur in many other decision contexts. Experiments on person memory and scripts have demonstrated the influence of expectations on memory for and inferences about people and events. For example, Cantor and Mischel (1977) presented participants with trait descriptions of four fictional people who were described as either extroverts or introverts. Participants later completed a recognition test on the descriptions, rating how certain they were that they had seen each test item. Some test statements included previously unseen traits that were conceptually related to the general trait label "extrovert" or "introvert"; other statements included new traits that were unrelated to the general trait label. Participants were more confident that they had seen new items that were conceptually related than conceptually unrelated to the general label. This suggests that the general trait label "extrovert" or "introvert" created an expectancy for conceptually related information.
Similar results were obtained using a face recognition paradigm (Locksley, Stangor, Hepburn, Grosovsky, & Hochstrasser, 1984). Participants watched 324 slides of faces of people who had allegedly scored as extroverts on a personality test. They were instructed to remember the faces because they would be tested on them later. Thirty of the faces were rated by pretest subjects as highly extroverted (expectancy congruent), and thirty faces were rated as highly introverted (expectancy incongruent). After viewing the acquisition series, participants were shown another set of 120 faces, 60 of which had appeared in the previous series (30 extroverted and 30 introverted), 60 of which were new (30 extroverted and 30 introverted). Participants judged whether each face was old (previously seen) or new, and their hit and false alarm rates were computed. The hit rate was higher for expectancy-congruent than for expectancy-incongruent faces. Consistent with the Cantor and Mischel (1977) study, the false alarm rate was higher for congruent than incongruent faces. Thus, the expectancy created by the trait label influenced participants' intrusion errors, with new congruent faces being falsely recognized as old more often than new incongruent faces. A higher false alarm rate for expectancy-congruent than expectancy-incongruent information has been obtained across a variety of experimental materials, including behavioral descriptions of social groups (Woll & Graesser, 1982), personality types (Cantor & Mischel, 1977; Locksley et al., 1984; Woll & Graesser, 1982), and script actions (Graesser, Gordon, & Sawyer, 1979; Graesser, Woll, Kowalski, & D. A. Smith, 1980; D. A. Smith & Graesser, 1981).

Given the generality of these expectancy effects across decision contexts, we predicted that people's prior knowledge of crime categories would also operate as an expectancy for information processing. Two potential manifestations of this expectancy role were explored. First, people should be more willing to accept new information about an event when that information is expectancy-congruent (typical of the crime) than expectancy-incongruent. In Study 1, we described a criminal event, then provided participants with supplementary information about the event that was either typical of the crime (expectancy-congruent) or atypical of the crime (expectancy-incongruent). We then assessed participants' memories for the original event to determine whether typical supplementary information was reported as fact more often than atypical supplementary information. Second, even in the absence of supplementary information, people should be more willing to fill gaps in the evidence with typical (expectancy-congruent) than atypical (expectancy-incongruent) information. If so, their intrusion errors will be biased toward typical information. Studies 2 and 3 tested this hypothesis.

**STUDY 1**

If people's prior knowledge of crime categories operates as an expectancy for information processing, then it should influence their acceptance of new information about an event. The goal of Study 1 was to investigate whether people are more willing to incorporate new information into a memory report of a criminal event when that information is expectancy-congruent (typical of the crime) than expectancy-incongruent (atypical of the crime). Considerable research has demonstrated that a person's memory report of a criminal event can be altered by pre-
senting postevent information; false information contained in misleading questions is often reported later as part of the original event (e.g., Belli, 1989; Loftus, 1975, 1977; Loftus, Miller, & Burns, 1978; V. L. Smith & Ellsworth, 1987; Tversky & Tuchin, 1989). In one experiment, participants watched a videotape of a traffic accident and then answered questions about it (Loftus, 1975). Half of the participants were asked, “How fast was the white sports car going when it passed the barn while traveling along the country road?” This question is misleading because there was no barn in the videotape. The other half of the participants were asked an unbiased version of the same question; “How fast was the white sports car going while traveling along the country road?” When later asked if they had seen a barn, 17% of the participants who heard the misleading question reported seeing a barn whereas only 3% of the participants who heard the unbiased question made that error. Thus, the postevent presentation of misleading information can impair the accuracy of people’s memory reports of the original event.

This misleading question paradigm provides a useful model for testing our hypothesis that typical information will be more readily incorporated into a memory report than atypical information. Specifically, if prior knowledge of crime categories operates as an expectancy for information processing, it should be easier to mislead people with false information that is typical of the crime than false information that is atypical. Typical misinformation is expectancy-congruent and should be easy to assimilate into one’s representation of the event. Atypical misinformation is expectancy-incongruent and should be relatively easy to reject. To test this hypothesis, we varied the typicality of the misleading information presented and later measured the frequency with which participants incorporated the misleading information into their memory reports.

**Method**

**Overview**

Participants listened to an audiotaped description of a burglary (Stage 1). After a 10-min filler task, they completed a questionnaire containing several critical questions (Stage 2). Some questions contained misleading information that is typical of burglary, some contained misleading information that is atypical of burglary, and some contained no misleading information. After another 10-min filler task, participants completed a questionnaire measuring the effects of the misleading information on their memory reports about the crime (Stage 3).

**Participants**

Participants were 91 introductory psychology students at a Midwestern research university. They participated for course credit in groups of two to five people.
**Materials and Procedure**

**Pretesting.** Pretesting was undertaken to determine what information is perceived by participants as typical and atypical of burglary. Twenty pretest participants listened to an audiotaped narrative of a burglary, which lasted about 4.5 min. The narrative began by describing the problems Steve, the perpetrator, is having at work and his fear that he will lose his job. To relieve stress, Steve frequently drives through a posh neighborhood in which one of his former college classmates, Dan McCormick, lives. Steve recalls in detail the expensive furnishings and electronic equipment Dan's parents have in their home. As he approaches the McCormicks' house, Steve notices that the owners are apparently away on vacation. Steve lets himself in and tours the house trying to decide what to take. The story ends with Steve loading some of the McCormicks' belongings into his car and leaving the premises.

After listening to the audiotape, the pretest participants rated how typical, or characteristic, of the crime several new details would be. For example, these new details included: “Steve has a high-paying job,” “Steve has a low-paying job,” “the crime occurs at night,” “the crime occurs during the day.” Participants rated the typicality of each detail on a 15-point scale ranging from -7 (very atypical), through 0 (neither typical nor atypical), to +7 (very typical). Different instantiations of the same feature (type of job, time of day) were included so that both typical and atypical versions of the feature could be identified. Features were chosen for Study 1 if one version received a negative average rating, another version received a positive average rating, and at least one of the averages was significantly different from zero. In this way, seven features were selected: (1) the time of day the crime occurred—typical = daytime ($M = 2.05$, different from zero at $p < .05$), atypical = nighttime ($M = -1.40$, n.s.); (2) the type of jacket Steve wore—typical = dark jacket ($M = 2.65$, $p < .001$), atypical = suit jacket ($M = -2.60$, $p < .001$); (3) the value of the stolen property—typical = high value ($M = 5.40$, $p < .001$), atypical = low value ($M = -3.70$, $p < .001$); (4) Steve's social class—typical = low ($M = 1.50$, $p < .05$), atypical = high ($M = -4.15$, $p < .001$); (5) Steve's knowledge of an alarm—typical = no knowledge ($M = 2.10$, $p < .01$), atypical = knowledge ($M = -2.35$, $p < .01$); (6) status of Steve's job—typical = low-paying ($M = 4.50$, $p < .001$), atypical = high-paying ($M = -4.80$, $p < .001$); (7) damage done to McCormicks' house—typical = no damage ($M = 4.15$, $p < .001$), atypical = damage ($M = -4.40$, $p < .001$). These typical and atypical details were used to construct the questionnaires for Study 1.

**Study 1.** In Stage 1 of this experiment, participants listened to the audiotaped narrative about Steve. They then completed a 10-min filler task unrelated to the experiment, in which they rated their familiarity with vocabulary words. In Stage 2, participants answered 16 questions about the burglary, 7 of which were critical for the experiment. Some critical questions contained misleading information, others contained only unbiased information. Based on the pretest ratings, three versions of each critical question were prepared for the Stage 2 questionnaire—one containing a new detail that is typical of burglary, one containing an atypical new detail, and one containing no new detail. The new details constituted misleading
information because they were not contained in the narrative. For example, the audiotape described Steve parking his car in the McCormicks' driveway, putting on his jacket, and looking around to see whether any neighbors were watching him. Each participant answered one of the following questions: (1) As Steve got out of his car and put on his dark jacket, did he look around for any neighbors who might be watching him? (dark jacket is typical misleading information); (2) As Steve got out of his car and put on his suit jacket, did he look around for any neighbors who might be watching him? (suit jacket is atypical misleading information); or (3) As Steve got out of his car and put on his jacket, did he look around for any neighbors who might be watching him? (unspecified jacket type, unbiased information).

There were three forms of the Stage 2 questionnaire which contained the different versions of the critical questions. For example, form 2 included typical misleading information about Steve's social class and his knowledge of an alarm system, atypical misleading information about the type of jacket Steve wore and damage done to the McCormicks' house, and unbiased information about the time of day the crime occurred, the value of the stolen property, and the status of Steve's job. The different versions of each critical question were randomly assigned to the three forms of the questionnaire. Thus, each participant's questionnaire contained two or three typical misleading questions, two or three atypical misleading questions, and one to three unbiased questions.

After the Stage 2 questionnaire, participants completed another unrelated 10-min filler task rating vocabulary words. Then, in Stage 3, they answered more questions about the burglary. The Stage 3 questionnaire was designed to test the effect on participants' memory reports of the misleading information presented in Stage 2. There were 22 questions, 7 of which probed for the critical items. The response options for each question included the typical misleading information, the atypical misleading information, and "was not mentioned." For example, one question asked: "What type of jacket was Steve wearing?" The response options were: "suit jacket," "dark jacket," and "was not mentioned." The typical response option was listed first for some questions; the atypical option was listed first for others. Participants were instructed to answer the questions based on the facts presented in the story, and not to make any assumptions about things that were not explicitly mentioned. If the story did not provide the answer to a question, they should circle the "was not mentioned" option. For the critical questions, "was not mentioned" was the correct answer. To ensure that the critical questions were not unusual, "was not mentioned" was the correct answer for some of the filler questions as well. For the other filler questions, one of the substantive response options was correct. After the experimental session, all participants were fully debriefed.

**Results and Discussion**

Each participant answered questions in Stage 2 containing typical misleading, atypical misleading, and unbiased features. For each feature type, we computed the proportion of questions in Stage 3 for which the participant endorsed the typical response option, the atypical response option, and "was not mentioned" (which was the
correct answer). The average proportions across participants are displayed in Table 1. The proportions were submitted to a Feature Type (typical misleading, atypical misleading, unbiased) \( \times \) Response Type (typical, atypical, correct) repeated measures ANOVA. This analysis yielded a main effect of Response Type, \( F(2,180) = 180.0, p < .0001 \), which was qualified by a significant interaction, \( F(4,360) = 28.4, p < .0001 \). The interaction indicates that the distribution of responses differed across feature types.

To explore the interaction, separate ANOVAs were conducted comparing the response distributions for typical misleading with unbiased features and atypical misleading with unbiased features. Both analyses yielded significant interactions: typical misleading versus unbiased, \( F(2,180) = 124.3, p < .0001 \); atypical misleading versus unbiased, \( F(2,180) = 3.52, p < .05 \). Thus, relative to the unbiased features, the presentation of both typical misleading and atypical misleading information in Stage 2 did influence participants' Stage 3 responses. Additional analyses were conducted to explore the nature of this influence. We predicted that typical misinformation would produce more endorsements of the typical response option and atypical misinformation would produce more endorsements of the atypical response option. A matched-pairs \( t \)-test revealed that the proportion of typical responses after hearing typical misinformation (.50) was significantly greater than the proportion of typical responses after hearing unbiased information (.22), \( t(90) = 6.80, p < .0001 \). Thus, participants were misled by the typical misinformation. Similarly, the proportion of atypical responses after hearing atypical misinformation (.09) was significantly greater than the proportion of atypical responses after hearing unbiased information (.01), \( t(90) = 4.14, p < .0001 \). Thus, participants were also misled by the presentation of atypical misinformation. Which type of misinformation was most influential? As predicted, participants were misled significantly more often by typical (.50 - .22 = .28) than by atypical misinformation (.09 - .01 = .08), \( t(90) = 4.13, p = .0001 \).

Table 1. Mean Proportions of Typical, Atypical, and Correct Responses (in Stage 3) for Questions that Contained Typical Misleading, Atypical Misleading, and Unbiased Features (in Stage 2)

<table>
<thead>
<tr>
<th>Response option</th>
<th>Features presented in Stage 2</th>
<th>Typical misleading</th>
<th>Atypical misleading</th>
<th>Unbiased</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical participants endorsed in Stage 3</td>
<td>.50</td>
<td>.20</td>
<td>.22</td>
<td></td>
</tr>
<tr>
<td>Atypical</td>
<td>.02</td>
<td>.09</td>
<td>.01</td>
<td></td>
</tr>
<tr>
<td>Correct</td>
<td>.47</td>
<td>.71</td>
<td>.78</td>
<td></td>
</tr>
</tbody>
</table>

4Misleading postevent information studies usually present analyses of participants' error rates on the critical questions. Because participants in this experiment could make two kinds of errors, endorsing either the typical or the atypical response option, the analyses of the response distributions reported in the text are more informative. However, the pattern of results is the same for analyses of error rates; participants made more errors (collapsed across error type) when they were presented with either typical or atypical misinformation than when they saw only unbiased information, and typical misinformation was more influential than atypical misinformation.
In this experiment, the presentation of both typical and atypical misleading information significantly influenced participants’ memory reports. Furthermore, as predicted, participants were significantly more influenced by typical than atypical misinformation. This finding provides a new perspective on what makes a misleading question persuasive. Previous research revealed that the impact of misinformation depends in part on the central or peripheral nature of the item in question (Dritsas & Hamilton [cited in Loftus, 1979]) and on the expertise of the source of the misinformation (V. L. Smith & Ellsworth, 1987). The current experiment reveals that the nature of the misinformation itself also influences its power; participants were misled significantly more often by typical than atypical items.

We proposed that if people’s prior knowledge of a crime category operates as an expectancy for information processing, they would be more likely to report as fact misinformation that is consistent with their prior knowledge (expectancy-congruent) than misinformation that is inconsistent (expectancy-incongruent). The results of Study 1 support this hypothesis. A second means of testing the expectancy role of people’s prior knowledge is to consider memory intrusion errors. If prior knowledge of a crime category operates as an expectancy, then intrusions should more frequently contain typical than atypical information. This finding would indicate that prior knowledge has a more powerful influence on memory reports than demonstrated in Study 1. Not only are people willing to accept typical misinformation asserted by a knowledgeable questioner, but they will also fill gaps in their stories with typical information on their own initiative. The distribution of responses for the unbiased questions in Study 1 provides some preliminary evidence that this hypothesis is correct. First, intrusion errors do occur—23% of the responses were errors endorsing features that were not specified in the story. Second, there appears to be a strong preference for typical intrusions; 96% of the intrusion errors were endorsements of typical information. Notice also that participants who heard atypical misleading information made more errors endorsing the typical (.20) than the atypical (.09) response option. This suggests that intrusions of typical information are a powerful source of error. It is possible, however, that the frequency or nature of these intrusions was influenced by participants’ exposure to misleading questions. These questions may have encouraged participants to fill gaps in the evidence because some gaps were filled for them. Study 2 was a preliminary investigation of whether people fill gaps in the available evidence without such encouragement, and, if so, whether these intrusions contain typical or atypical information.

STUDY 2

This study explored the frequency and typicality of memory intrusion errors. Participants listened to the audiotaped burglary from Study 1, did an unrelated 10-min filler task, then completed the Stage 3 questionnaire, which assessed the accuracy of their memory reports for the critical items. They did not complete Stage 2 of the previous experiment, and therefore did not receive any misleading information.
Method

Participants

Participants were 18 introductory psychology students at a Midwestern research university. They participated for course credit in groups of up to five people. The goal of Study 2 was to ascertain whether participants make unprompted intrusion errors in crime stories, and whether there is a bias toward intrusions of "typical" information. Because of the preliminary nature of this study, a total of 18 participants was considered ample.

Procedure

Participants listened to the audiotape of the burglary from Study 1. After an unrelated 10-min filler task, they completed the Stage 3 questionnaire from Study 1. This questionnaire contained seven critical questions which assessed participants’ memories for facts that were left unspecified on the audiotape. The response options included a typical answer, an atypical answer, and a "was not mentioned" option (which was the correct answer). Two additional response options were added for this study: “correct answer not available,” and “don’t remember.” The former provided an option for participants who filled the gap in their memory with a specific detail that did not correspond to the particular typical or atypical answers provided. The latter allowed participants to acknowledge they did not know the answer, and was included so that they were not forced to make intrusion errors. Participants were instructed to base their answers solely on the audiotape they heard. They should circle the “was not mentioned” option if the answer to the question was not explicitly stated on the audiotape, the “correct answer not available” option if the answer was provided on the audiotape but was not included among the available response options, and the “don’t remember” option if they could not remember whether the information was provided on the audiotape. After completing the study, all participants were fully debriefed.

Endorsements of the typical, atypical or “correct answer not available” options all represent intrusions. Participants have filled a gap in the story with specific information about the feature in question. If participants’ prior knowledge of burglary functions as an expectancy for information processing, their intrusions should contain typical information more often than atypical information. Thus, there should be a systematic tendency to endorse the typical answer.

Results and Discussion

Of the 18 participants in this study, 17 (94%) made intrusion errors. For each participant, we computed the percentage of questions (out of seven) for which each response type was endorsed. On average, 56% of participants’ responses were correct (“was not mentioned”) and 5% were “don’t remember.” The remaining 39%
of responses were intrusion errors, indicating that participants filled gaps in the available information quite frequently. Typical intrusions constituted 32% of the total, atypical intrusions 1%, and "correct answer not available" 6%. As predicted, typical intrusions (32%) were significantly more frequent than atypical intrusions (1%), \( t(17) = 7.38, p < .0001 \). We did not know what information participants had in mind when they endorsed the "correct answer not available" option, so we could not determine whether these intrusions contained typical or atypical information. However, even if all of the unspecified intrusions were atypical, there would still be a strong preference for typical intrusions (32% versus 7%), \( t(17) = 5.42, p < .0001 \). This study provides preliminary evidence that participants do fill gaps in the evidence without explicit suggestion to do so, and that they tend to fill those gaps with typical information. This finding encouraged us to conduct a more rigorous test of our intrusion hypothesis.

Gaps may occur in a person's memory of an event in two ways. First, no information about the item in question is presented. Second, information is provided, but subsequently forgotten. If people's prior knowledge operates as an expectancy, both types of gaps should be filled more frequently with typical than atypical information. Study 3 was designed to test this expanded hypothesis.

In Study 3 we used a yes–no response format. This afforded several advantages over the dependent measure used in Study 2. First, participants were no longer forced to choose between the typical and atypical response options. Participants may prefer the typical over the atypical answer when both are presented simultaneously, but be quite willing to endorse the atypical answer in the absence of this forced choice. The yes–no format permitted participants to evaluate each item in isolation, and prevented any artificial suppression of the atypical response which may have occurred due to the typical–atypical forced choice format used in Study 2. Second, the collection of response options that participants faced in Study 2 was rather unusual; some answers were specific, others general. This may have produced some confusion over the meaning of the response options. The yes–no response format used in Study 3 simplified the task to prevent such confusion. Third, in Study 3 we were able to measure both the hit and false alarm rates for typical and atypical information.

**STUDY 3**

**Method**

**Overview**

Participants listened to an audiotaped description of a burglary that contained some features that are typical of burglary, some features that are atypical of burglary, and some features for which no specific details were given. After an unrelated 10-min filler task, participants took a memory test on the story.
**Participants**

Participants were 74 introductory psychology students at a Midwestern research university. They participated for course credit in groups of two to five people.

**Procedure**

Participants listened to an audiotape of the burglary story from Study 1, which was modified for this experiment. In the previous studies, all of the critical features were left unspecified in the story; no information was provided about them. In this experiment, specific information was given about some of the critical items. Of the six critical items used in this experiment, two provided typical information, two provided atypical information, and two were left unspecified. Three versions of the story were then developed; for each version it was randomly determined which features were typical, atypical, and unspecified, with the constraint that each feature appear once in typical form, once in atypical form, and once in unspecified form across the three versions.

After listening to the audiotape, participants completed a 10-min filler task unrelated to the experiment. They then listened to a series of statements about the burglary that was presented on audiotape. There were two statements for each critical feature. One contained the typical information, the other the atypical information. For example, “Steve was wearing a dark jacket at the time of the crime” contains typical information about jacket type, and “Steve was wearing a suit jacket at the time of the crime” contains atypical information about jacket type. The 12 critical statements and 30 filler statements were presented in random order. Participants were instructed to indicate for each statement whether the information conveyed was mentioned in the story. They responded on a 4-point scale, labeled -2 (definitely not mentioned in story), -1 (possibly not mentioned in story), +1 (possibly mentioned in story), +2 (definitely mentioned in story). This scale provided a measure of the accuracy of participants' memories for the critical items and a crude measure of their confidence in those memories. At the conclusion of the experimental session, all participants were fully debriefed.

**Results**

Participants responded to one typical and one atypical statement about each of two typical, two atypical, and two unspecified story items (12 statements of 6 types). Analyses were conducted both on the raw scores from the 4-point scale and on a dichotomized recoding of the scale, where scores of +1 and +2 indicated that the information was mentioned in the story (“yes”) and -1 and -2 indicated that the information was not mentioned in the story (“no”). Because the analyses of raw scores and proportions yielded identical results, only the proportion data will be reported. Table 2 shows the mean proportion of yes responses to typical and atypical statements in each information condition.
We predicted that there would be more intrusions of, or false alarms to, typical than atypical information. Two comparisons are relevant. First, gaps in memory may occur because the information provided in the story is incomplete. In this situation, participants should be more willing to fill gaps with typical than atypical information. As predicted, when features were left unspecified in the story, typical false alarms (.40) were significantly more frequent than atypical false alarms (.12), $F(1,73) = 28.6, p < .0001$. Second, gaps in memory may also result from forgetting information that was provided in the story. In this situation as well, participants should be more willing to fill the gaps with typical than atypical information. So, false alarms should be more frequent when participants are provided atypical information in the story and respond to a statement containing typical information than when they hear typical information in the story and respond to a statement containing atypical information. As predicted, false alarms were significantly more frequent on typical (.20) than atypical (.11) statements, $F(1,73) = 4.54, p < .05$. Hit rates for typical (.87) and atypical (.85) story items were equally high, indicating that participants’ memories for the story details were generally quite good, $F < 1$.

**GENERAL DISCUSSION**

These studies indicate that people’s prior knowledge of crime categories does operate as an expectancy for information processing. Study 1 demonstrated that people’s memory reports are more susceptible to influence by typical than atypical misleading information. Thus, people can be persuaded to fill gaps in the evidence and are more likely to do so with expectancy-congruent than expectancy-incongruent information. All three studies, using different procedures and measures, revealed that memory intrusions are also more likely to contain typical than atypical information. This indicates that people are willing to fill gaps in the evidence even without explicit suggestion by an outside source, and that they tend to fill those gaps with expectancy-congruent information.\(^5\)

\(^5\)Consistent with these conclusions, another recent study found that unstated items from a robbery script were incorrectly recognized as “old” significantly more often than unstated nonscript items (Holst & Pezdek, 1992).
These findings reveal that people's prior knowledge of crime categories plays a broader role than previously appreciated. Earlier research demonstrated its influence on the verdict-selection process (V. L. Smith, 1991, 1993); the current research implicates it in the fact-finding process as well. People's willingness to use their prior knowledge of crime categories for fact-finding may have important implications for trial outcome. If jurors fill gaps in the available evidence with typical information, then the resulting representation of the event may be perceived as more typical of the crime charged than if such gap-filling had not occurred. Witnesses testifying at trial may also fall prey to their expectancies, filling gaps in their memories with typical information. Thus, the accepted misinformation or self-generated intrusions of both witnesses and jurors may operate to artificially inflate the perceived typicality of the event. Because typicality judgments influence verdict choice, these processes may result in a higher probability of a guilty verdict than is warranted by the evidence presented against the defendant.

People's prior knowledge of crime categories may influence the fact-finding process in other ways as well. First, people may better remember typical than atypical information. Study 3 did not show such an effect—the hit rates were the same for typical and atypical information. However, the retention interval used in this experiment was quite short, only 10 min. Given more time to decay, people's memories may well favor typical over atypical information (Stangor & McMillan, 1992). Second, jurors may interpret ambiguous information as expectancy-congruent, or third, they may resolve disputed issues in favor of typical information. Thus, there may be many ways in which people's prior knowledge of crime categories serves as a framework for information processing. Future research should explore these possibilities.

It is also important to consider the limitations of this research. The studies reported in this article indicate that people's prior knowledge of crime categories operates as an expectancy for information processing. However, there are several differences between the tasks confronting participants in these studies and the task confronting jurors in a real criminal trial. First, participants in these studies heard a short description of the details of a burglary told in narrative form. This is quite different from the usually large, complicated, and often conflicting body of evidence presented in actual trials in a question-answer format. Because this research explored the role of people's prior knowledge of crime categories on the fact-finding process, it was important that all participants base their responses on the same information. To this end, the details were presented in a straightforward, coherent, and internally consistent manner; participants did not have to interpret or make inferences about the evidence to generate a plausible representation of the event. This simplification minimized variability in participants' performance due to differing interpretations of the evidence or differing judgments of witness credibility. Having established that people's prior knowledge of crime categories can have powerful effects on their processing of available evidence, future research should replicate these effects using more realistic stimulus trials. We are encouraged to predict that more complex stimuli would not materially alter the effects we report here. Previous research on people's prior knowledge demonstrated powerful typicality effects using only 5–6-sentence descriptions of a crime. The burglary scenario used in the current
Smith and Studebaker

studies was considerably longer and more complicated than the stimuli previously used and we still find strong typicality effects. Thus, we have no reason to expect that further increases in complexity would change these effects, though clearly this matter must ultimately be resolved by future research.

Second, other task characteristics differed for our participants from real trials. For example, the retention intervals were quite short, only 10 min. Real jurors face retention intervals on the order of hours, days, or weeks. Nevertheless, even with short retention intervals we obtained clear typicality effects. Our goal was to produce less than perfect memory performance in order to test our hypotheses, and a short retention interval was sufficient to achieve that goal. Longer retention intervals would have allowed for greater memory decay, but were not necessary to obtain the effects we were testing. Our dependent measures were presented in multiple-choice and true–false formats, whereas jurors rely principally on recall. We chose the recognition format for these studies to facilitate the classification and quantification of participants’ responses. Previous research has obtained typicality effects on verdict choice using both recall and recognition measures (V. L. Smith, 1991), so we felt the gain in interpretability afforded by recognition measures justified the sacrifice of mundane realism. Furthermore, although jurors may rely primarily on recall, recognition probably also plays a role in a jury’s discussion of the evidence. Disputes between jurors may present conflicting pieces of information and the jury must choose which version is an accurate representation of the evidence or testimony. In this situation the task is more like a recognition task—which piece of information is correct? So, although jurors may rely more heavily on recall, a recognition task is not beyond the bounds of their experiences.

Another departure from actual jury trials was the subject population used in these studies. All participants were undergraduates at a Midwestern research university, who are younger and better educated than the average juror. It is possible that on these tasks the performance of actual jurors might differ from the performance of students. If such differences exist, one might predict that students would perform more accurately. Students are more practiced at taking tests, may be more practiced at setting aside their preconceived notions, and may be more adept at resisting social influence attempts. Nevertheless, these studies demonstrated that the students were susceptible to misleading postevent information and did make intrusion errors. It is doubtful whether a more diverse population of jurors would avoid such errors. In the end, of course, all of these generalizability issues raise additional testable hypotheses. Whether the effects obtained in the studies reported in this article are robust across these variations in task characteristics and subject population is an important topic for future research.

The discovery that people’s prior knowledge of crime categories influences fact-finding raises important questions about the appropriateness of such influence. Should people use their prior knowledge as a basis for making inferences about the evidence? The simple answer to this question is no—we do not want jurors filling in details of a particular defendant’s crime based on “what burglaries typically look like.” However, jurors are encouraged to use their prior knowledge of the world as a guide for fact-finding. At what point do a person’s prior knowledge of crime categories and prior knowledge of the world diverge? Suppose, for example,
there is a dispute at trial over whether the perpetrator of a robbery was armed. Suppose further that one juror concludes that the perpetrator was armed because most people who make threats have the means to back them up and another juror concludes that the perpetrator was armed because weapons are typical features of robberies. The first juror's conclusion is based on an assumption about the social world and the second juror's conclusion is based on a prototypic representation of the crime category. However, we do not discern a clear difference between these two inferences; both are based on jurors' beliefs about what is typical in such situations. Therefore, there may be some circumstances in which people's prior knowledge of crime categories is properly used as a guide for fact-finding and other circumstances in which such influence is inappropriate. Discussion and debate within the legal and psychological communities over how to distinguish proper from improper uses of prior knowledge is a promising way of addressing such fundamental questions about how jurors should and should not make decisions. Developing interventions that will reduce or eliminate the unwanted influences of prior knowledge is then an important topic for future research.

REFERENCES


LAY PERCEPTIONS OF JUSTICE VS. CRIMINAL LAW DOCTRINE: A FALSE DICHOTOMY?

Dan M. Kahan*

It is nearly impossible to overstate the importance of Robinson and Darley’s work on the relationship between lay perceptions of justice and doctrines of substantive criminal law. Consequently, it is nearly impossible to advance an important criticism of it. I do, however, believe that there is one significant gap in their methodology that they ought to fill as their project continues to unfold.

This gap is their failure to address the importance of prototypical reasoning by jurors. A growing body of experimental studies suggests that jurors are not very much influenced by formal tests embodied in the substantive criminal law doctrines; instead they consult prototypical representations, absorbed from their immersion in social and cultural life, to determine whether the facts add up to a particular crime or defense. Because jurors seem more or less automatically to understand the law to be whatever they intuitively think it is, Robinson and Darley may well be overstating the existence of conflict between lay sensibilities and law.

The work I have in mind is in fact an offshoot of work in cognitive psychology that tries to make sense of the phenomenon of intuitive perception—our capacity to “know” something when we “see” it, without being able to explain why. For example, we distinguish among different colors, different sounds, and different tastes without being able to articulate the bases on which we are making such discriminations. Each of us learns to recognize thousands of faces, a skill that even the most power-

* Professor of Law, Yale Law School.
3. See id. at 870.
ful computers cannot be programmed to emulate. Whether grammar can be reduced to a set of stateable rules—an issue on which philosophers of language disagree—mastery of a language clearly does not depend on being aware of such rules or on consulting them before one speaks. In all of these settings and in countless others, we make critical judgments, less by reflectively deducing what is important than by intuitively apprehending it.

Psychologists link intuitive perception to a cognitive operation that they refer to as pattern recognition. According to this theory, what goes on in the brain when individuals recognize faces, construct grammatical sentences, play chess, or perceive tastes, colors, and smells is not a form of algorithmic computation, but rather a rapid, pre-verbal cycling process whereby the case at hand is compared to, contrasted with, and ultimately matched against a wide range of mentally inventoried prototypes.

Experimental data suggests that moral and other types of evaluative reasoning fit this profile. Individuals intuitively “recognize cruelty and kindness, avarice and generosity, treachery and honor, mendacity and honesty, the cowardly way out and right thing to do” by drawing on “a hierarchy of moral prototypes,” compiled from their exposure to “a substantial number of relevant examples of the moral kinds at issue.”

How about legal concepts? They have also been shown to display the key features of pattern recognition.

Consider the insanity defense. Mock jury studies show that jurors instructed under the liberal Model Penal Code (“MPC”) test, which excuses both volitional and cognitive defects, are no more likely to convict or acquit than those instructed under the traditional M’Naghten rule, which focuses only on cognitive impairments; indeed, the verdicts of

6. See Churchland, supra note 4, at 143-44.
7. Many other examples involve specialized tasks. Reading x-rays, interpreting aerial photographs, and grading of commodities such as cheese and wool all involve perceptual intuition. See Eleanor J. Gibson, Principles of Perceptual Learning and Development 6-9 (1969). Expert play in chess also involves perceptual intuition. See Margolis, supra note 5, at 104-05.
8. See Margolis, supra note 5, at 1.
9. See id.; Churchland, supra note 4, at 21-34.
10. Churchland, supra note 4, at 144, 146; see also Margolis, supra note 5, ch. 5 (discussing how people check judgments and noting the role of social acceptance); Gerald Dworkin, Unprincipled Ethics, 20 MIDWEST STUD. PHIL. 224 (1995) (discussing the nature of moral judgments).
12. See James R.P. Ogloff, A Comparison of Insanity Defense Standards on Juror De-
those instructed under either test are indistinguishable from those given an “insanity” instruction that specifies no definition of the term. However instructed, juries determine the offender’s sanity according to a lay construct that focuses on a wide array of extra-doctrinal considerations, including “the defendant’s background,” his “relationship with the victim,” his “intent to harm,” and his “culpability” before the act for bringing about [his] incapacity.” Just as significant, jurors’ lay construct of insanity shifts unpredictably as experimenters alter the facts. The considerations that strike jurors as decisive in one case turn out to be irrelevant, or decidedly less important than some other previously irrelevant consideration, in the next case.

Consequently, although it is possible to specify the factors that juries take into account, it is not possible to systematize those factors into a stateable rule or test. Like other phenomena involving pattern recognition—from the identification of faces, to the construction of grammatical sentences, to the discernment of anger or fear in others—the perception of insanity is a task that individuals can perform but not explain in a verbally cogent manner.

The same goes for jurors’ perception of various types of criminal offenses. Experimental studies suggest that jurors identify crimes prototypically rather than algorithmically. That is, to determine whether a given set of facts constitutes “kidnapping,” “ murder,” “assault,” or “ robbery,” jurors consult inventoried prototypes, which consist not of necessary and sufficient conditions, but rather of collections of attributes, against which putative instances of a crime are judged more or less typical. Moreover, as with their perceptions of “insanity,” jurors’ perceptions of whether actions constitute particular crimes are unaffected by the definitions contained in the instructions that courts give them.

---

14. Id. at 526.
17. See Smith, supra note 2, at 868.
18. See id.
19. See id. at 870.
What is the significance of pattern recognition for Robinson and Darley’s work? The nerve of Justice, Liability, and Blame: Community Views and the Criminal Law is to document the convergence and, even more importantly, the divergence between formal doctrines and lay sensibilities of justice. But, if pattern recognition plays the role in juror decision-making that this work suggests, their anxieties about the divergence may well be overstated.

Substantive criminal law doctrines—from duress to insanity to self-defense—are highly abstract. To make these doctrines produce concrete results, experimental studies suggest that jurors draw on their pre-existing intuitions about what it is they are supposed to be looking for. And if doctrine is given content through lay sensibilities in this way, there is no reason to expect conflict between laws and sensibilities.

Robinson and Darley, of course, do purport to find conflicts of this very sort. But the conflict may be more apparent than real. Having been socialized to apprehend the theoretical aims of the drafters of the MPC, Robinson and Darley are able to say when the MPC’s “substantial step” test, for example, would produce results different from the common-law “dangerous proximity” test. But jurors have not been seeped in those theories. They have their own prototypes, drawn from a wide-range of everyday sources of social learning, about what counts as an “attempt.” Their sensibilities of what counts as an attempt might not conform to what the MPC regards as one. But by virtue of the influence that pattern recognition exerts over legal decision-making, there is no reason to believe that jurors will apply the MPC test, the common-law test, or any other abstract doctrine in a way that disappoints their sensibilities.

At a minimum, the existing experimental work on prototypical reasoning and juror decision-making suggests that Robinson and Darley should modify their methodology in an important way. Whenever they conclude that there is in fact a divergence between lay sensibilities and the results that they believe doctrine entails, they should run an additional test to see whether mock jurors, supplied with the relevant doctrine, would in fact apply the doctrine in a way that disappoints their sensibilities. Consistent with existing experimental work, my hypothesis is that they will rarely, if ever, see this result.

20. See Robinson & Darley, supra note 1, at xv-xvi.
21. See Smith, supra note 2, at 870.
22. See Robinson & Darley, supra note 1, at 14-23.
23. See id., at 205.