

2013

Using a Prediction and Option Generation Paradigm to Understand Decision Making

Joel Suss
Michigan Technological University

Follow this and additional works at: <https://digitalcommons.mtu.edu/etds>



Part of the [Cognitive Psychology Commons](#)

Copyright 2013 Joel Suss

Recommended Citation

Suss, Joel, "Using a Prediction and Option Generation Paradigm to Understand Decision Making",
Dissertation, Michigan Technological University, 2013.
<https://digitalcommons.mtu.edu/etds/495>

Follow this and additional works at: <https://digitalcommons.mtu.edu/etds>



Part of the [Cognitive Psychology Commons](#)

USING A PREDICTION AND OPTION GENERATION PARADIGM TO
UNDERSTAND DECISION MAKING

By
Joel Suss

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Applied Cognitive Science and Human Factors

MICHIGAN TECHNOLOGICAL UNIVERSITY

2013

© 2013 Joel Suss

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Applied Cognitive Science and Human Factors.

Department of Cognitive and Learning Sciences

Dissertation Advisor: *Dr Paul Ward*

Committee Member: *Dr Edward Cokely*

Committee Member: *Dr Shane Mueller*

Committee Member: *Dr Neil Charness*

Committee Member: *Dr K. Anders Ericsson*

Department Chair: *Dr Bradley H. Baltensperger*

Table of Contents

List of Figures.....	vi
List of Tables.....	vii
Acknowledgements.....	ix
Abstract.....	xi
1. INTRODUCTION.....	1
Outline of literature review (chapters 2 & 3).....	4
2. LITERATURE REVIEW: SITUATION ASSESSMENT IN DECISION MAKING.....	8
Situation assessment in dynamic domains.....	8
Situation assessment in recognition-primed decision making.....	9
Situation assessment in the model of situation awareness.....	13
Situation assessment in other complex and dynamic domains.....	17
Hypothesis generation in ill-defined problems.....	23
Perceptual anticipation.....	28
Perceptual anticipation: Related constructs.....	29
Perceptual anticipation in complex domains.....	31
Perceptual anticipation in sport.....	32
Projection in dynamic domains.....	36
Hazard perception in driving.....	39
Theoretically-motivated studies of skill in perceptual anticipation.....	42
Summary.....	47
3. LITERATURE REVIEW: RESPONSE PHASE OF DECISION MAKING.....	48
Response generation and selection.....	48
Response generation/selection tasks to examine perceptual-cognitive expertise.....	49
Response option generation in ill-defined problems.....	53
Empirical studies of response option generation in complex and dynamic domains.....	56
Option generation/selection in the situation assessment and response phases on decision making.....	62
Summary.....	65
4. GENERAL HYPOTHESES AND OUTLINE OF STUDIES 1 AND 2.....	66
General hypotheses.....	66
Outline of Studies 1 and 2.....	67
5. STUDY 1.....	69
Study 1A.....	71
Method.....	72
Participants.....	72
Materials and stimuli.....	72

Option-generation tasks.....	74
Procedure.....	76
Data preparation.....	78
Study 1B.....	78
Cognitive task analysis.....	78
Method.....	78
Participants.....	78
Materials and stimuli.....	79
Procedure.....	79
Data analysis.....	80
Performance data.....	82
Method.....	82
Participants.....	82
Materials and stimuli.....	83
Procedure.....	83
Data preparation.....	85
Study 1C.....	85
Method.....	86
Participants.....	86
Materials and stimuli.....	86
Procedure.....	88
Data preparation.....	88
Data analysis and results for Studies 1A–C.....	89
Data screening.....	89
Anticipation scores.....	89
Analyses employed.....	89
Identification of discriminating scenarios.....	92
Effect size calculation.....	95
Summary of Studies 1A–C.....	95
6. STUDY 2.....	98
Hypotheses.....	99
Method.....	103
Power analysis.....	103
Participants.....	103
Materials and stimuli.....	104
Option-generation tasks.....	105
Assessment option-generation task.....	105
Response option-generation task.....	107
Post-option-generation cognitive task analysis.....	108
Retrospective verbal reports.....	109
Stimulated recall of in-event option generation.....	109
Procedure.....	111
Data analysis.....	112
Signal detection analysis.....	112
Dependent variables.....	113

Tests of the experimental hypotheses.....	118
Supplementary analyses.....	120
Cognitive task analysis data.....	121
Retrospective verbal report data.....	122
Stimulated recall data.....	122
Results.....	125
Signal detection analysis.....	125
Tests of the experimental hypotheses.....	126
Cognitive task analysis data.....	134
Retrospective verbal report data.....	137
Stimulated recall data.....	138
Quantitative analysis.....	138
Qualitative analysis.....	139
Summary.....	147
7. DISCUSSION.....	151
Outcome performance.....	153
Option-generation measures.....	155
Supplementary points.....	162
Signal detection performance.....	162
Taking the first-generated option.....	163
Entropy/information uncertainty.....	163
Limitations and future research directions.....	165
Summary.....	168
8. REFERENCES.....	170
9. APPENDICES.....	189
A. Scenarios used in Studies 1 and 2.....	190
B. Exemplar scenarios.....	191
C. Timeline of Blow Up scenario with description of critical cues and timing of occlusion points.....	193
D. List of occlusion point times (in seconds) for Studies 1A–C.....	194
E. Radio message/contextual introductions to the video scenarios.....	195
F. G*Power power analysis output.....	197
G. Thought statement types.....	198

List of Figures

- Figure 1. The mean number of task-relevant and task-irrelevant assessment/
response options generated, by experience group, for the six test trials..... 129
- Figure 2. Description of the phases of the Domestic Punch scenario..... 136

List of Tables

Table 1. Occlusion points selected by experienced police trainers in Study 1B (Cognitive task analysis).....	80
Table 2. Optimal occlusion point for discriminating scenarios, with results of classification criteria.....	94
Table 3. Effect size (Cohen's <i>h</i>) measures for the discriminating scenarios and the overall sample-weighted effect size.....	95
Table 4. Performance categories as functions of experience level, ability to predict the culminating event, and ability to take the best response option.....	124
Table 5. Proportion of times that participants drew/did not draw their gun by trial type and experience group.....	125
Table 6. Measures of sensitivity and bias by experience group.....	126
Table 7. Summary of multiple regression analysis for variables predicting the number of correct predictions made during assessment ($n = 49$).....	131
Table 8. Summary of multiple regression analysis for variables predicting the number of times participants took the best response option ($n = 49$).....	132
Table 9. Number of participants in each contrast group for the six test trials.....	135
Table 10. Central tendency measures of statement types elicited via retrospective verbal report (Domestic Punch scenario).....	137
Table 11. Median number of statements, by scenario phase (with duration, in seconds), and experience level (Exp. = Experienced, Less-exp. = Less-experienced) for the Domestic Punch Scenario.....	138
Table 12. Categorization of participants who completed the stimulated recall task for the Domestic Punch scenario.....	139
Table 13. Stimulated recall statements for Phase 1 (Confrontation) of the Domestic Punch scenario by statement type, success in responding, and experience group.....	141
Table 14. Stimulated recall statements for Phase 2 (Confrontation) of the Domestic Punch scenario by statement type, success in responding, and experience group.....	142

Table 15. Stimulated recall statements for Phase 3 (Confrontation) of the Domestic Punch scenario by statement type, success in responding, and experience group.....	143
Table 16. Inferences (underlined) and response options that highlight the similarity between experienced and less-experienced participants who took the best response.....	146
Table 17. Experience-based differences in response options reported by participants who did not take the best response.....	147

Acknowledgements

This dissertation would not have been possible without the support of many people. Many thanks to my adviser, Paul Ward, who read many revisions of this work, and provided much-needed feedback—especially when I could not see the forest for the trees. Thanks too, to the members of my dissertation committee, Edward Cokely, Shane Mueller, Neil Charness, and K. Anders Ericsson, whose guidance, assistance, and suggestions I greatly appreciate.

Thank you to the National Science Foundation, who provided support for this research in the form of a Doctoral Dissertation Research Improvement Grant. This grant provided the funds that allowed me to make numerous data-collection trips to locations across the United States of America.

Thank you to Robert McCue, General Manager, IES Interactive Training, Inc. for providing me with full and unrestricted access to IES' library of proprietary video scenarios.

Thank you to the following law enforcement researchers, who provided valuable input and advice, and helped me contact law enforcement agencies interested in collaborating on this research: Thomas Aveni, Executive Director, The Police Policy Studies Council; Dr J. Pete Blair, Assistant Professor, Texas State University – San Marcos; and Dr Terry Wollert, Senior Researcher, Federal Law Enforcement Training Center. An extra special thanks to Dr William Lewinski, Executive Director, Force Science Institute, who graciously allowed me to pick his brain, in person, for a whole day.

Thank you to the many law enforcement agencies and training academies that collaborated on this research by facilitating access to potential participants. Many agencies preferred to remain unnamed, and therefore, I elected not to name any. Nonetheless, I thank those who secured internal approval from their respective agencies so that I could conduct the research, and those that coordinated the testing sessions.

Thank you also to the law enforcement officers and recruits-in-training who volunteered to participate in the research. I truly admire the job that you do, and I appreciate the time you took out of your busy schedules to participate.

Finally, thank you to my parents, sister, and many friends who offered their support throughout the course of my studies.

Abstract

In many complex and dynamic domains, the ability to generate and then select the appropriate course of action is based on the decision maker's "reading" of the situation—in other words, their ability to assess the situation and predict how it will evolve over the next few seconds. Current theories regarding option generation during the situation assessment and response phases of decision making offer contrasting views on the cognitive mechanisms that support superior performance. The Recognition-Primed Decision-making model (RPD; Klein, 1989) and Take-The-First heuristic (TTF; Johnson & Raab, 2003) suggest that superior decisions are made by generating few options, and then selecting the first option as the final one. Long-Term Working Memory theory (LTWM; Ericsson & Kintsch, 1995), on the other hand, posits that skilled decision makers construct rich, detailed situation models, and that as a result, skilled performers should have the ability to generate more of the available task-relevant options.

The main goal of this dissertation was to use these theories about option generation as a way to further the understanding of how police officers anticipate a perpetrator's actions, and make decisions about how to respond, during dynamic law enforcement situations. An additional goal was to gather information that can be used, in the future, to design training based on the anticipation skills, decision strategies, and processes of experienced officers. Two studies were conducted to achieve these goals.

Study 1 identified video-based law enforcement scenarios that could be used to discriminate between experienced and less-experienced police officers, in terms of their ability to anticipate the outcome. The discriminating scenarios were used as the stimuli in Study 2; 23 experienced and 26 less-experienced police officers observed temporally-

occluded versions of the scenarios, and then completed assessment and response option-generation tasks.

The results provided mixed support for the nature of option generation in these situations. Consistent with RPD and TTF, participants typically selected the first-generated option as their final one, and did so during both the assessment and response phases of decision making. Consistent with LTWM theory, participants—regardless of experience level—generated more task-relevant assessment options than task-irrelevant options. However, an expected interaction between experience level and option-relevance was not observed.

Collectively, the two studies provide a deeper understanding of how police officers make decisions in dynamic situations. The methods developed and employed in the studies can be used to investigate anticipation and decision making in other critical domains (e.g., nursing, military). The results are discussed in relation to how they can inform future studies of option-generation performance, and how they could be applied to develop training for law enforcement officers.

CHAPTER 1

INTRODUCTION

When individuals make decisions in dynamic and complex environments, thinking about the situation and acting in pursuit of a goal are intimately linked (Orasanu, 1993). To understand skilled performance in highly representative or naturalistic settings, it is important to consider the relationship between the external environment, an agent's perception of that environment, and the actions taken to achieve an outcome (Cokely & Kelley, 2009; Smith & Hancock, 1995). Although the concept of a perception-action cycle is well established (e.g., Neisser, 1976), research on skilled performance in dynamic and complex environments has tended to focus on either the perception or the action phase of decision making (for state of the science reviews on expertise, see Ericsson, Charness, Feltovich, & Hoffman, 2006). For instance, there exists on the one hand, evidence that experienced performers are better at anticipating and assessing situations than their less-experienced counterparts with little reference to the courses of action taken (e.g., Calderwood, Crandall, & Baynes, 1990; Ward & Williams, 2003). On the other hand, complementary evidence exists that skilled performers generate and/or select better responses than less-skilled performers, however, the perceptual skills of these performers are rarely considered (e.g., Raab & Johnson, 2007). There is remarkably little research that has addressed, directly and empirically, the relationship between the perceptual-cognitive (i.e., situational assessment) and action (i.e., response generation/selection) phases of decision making in dynamic and complex environments. Therefore, the main goal of the current research is to further our understanding of the relationship between both phases, and to investigate the extent to which this relationship

is affected by skill level.

To achieve this goal, the current research will investigate situation assessment and response generation/selection behavior in complex, dynamic, time pressured, and potentially threatening law enforcement situations. The following account illustrates the importance of both phases of decision making in law enforcement:

. . . at dusk one day . . . my partner and I spotted three gang members walking down the sidewalk, facing away from us. As we rolled up . . . the subjects' heads start turning all over the place; it was evident they were looking for a place to run. The two guys on the ends split The guy in the middle didn't know which way to go, and he just froze there in front of us. We were right on top of him when we stopped—maybe ten feet away. As I was getting out of the passenger door of the car, he started digging in his waistband with his right hand. Then I could see that he was reaching into his crotch area, then that he was trying to reach toward his left thigh area, as if he was trying to grab something that was falling down his pants leg. He was starting to turn around toward me as he was fishing around in his pants. He was looking right at me and I was yelling at him not to move: "Stop! Don't move! Don't move! Don't move!" As I was giving him commands, I drew my revolver. When I got about five feet from the guy, he came up with a chrome .25 auto [handgun]. Then, as soon as his hand reached his center stomach area, he dropped the gun right on the sidewalk. We took him into custody, and that was that. (Klinger, 2004, p. 61)

As this account suggests, successful performance in law enforcement situations depends on an officer's ability to assess the situation accurately, and predict what will

happen next. Previous empirical research in this domain has found that experienced police officers took appropriate action (e.g., justifiably shot a gun-wielding suspect) significantly earlier than did police recruits, and suggested that successful performance was related to an officer's ability to anticipate situational outcomes (Harris, Tashman, Ward, Ericsson, & Eccles, 2006). Similarly, research in other complex and dynamic environments has found that experts make predictions frequently and successfully, often under temporal and/or informational constraint (e.g., McKenna & Crick, 1994; Sulistyawati, Wickens, & Chui, 2011; Williams, Ward, Knowles, & Smeeton, 2002).

The challenging nature of dynamic law enforcement situations is highlighted by the number of police officers killed and assaulted in the line of duty. Between 2002-2011, an average of 54 officers were killed feloniously each year in the United States of America (Federal Bureau of Investigation [FBI], 2012, Table 1). Using available data for 2009, police officers were killed by shooting at a rate of 6.4 per 100,000 officers (derived from FBI, 2010a and 2010b); this was substantially greater than the overall (i.e. total US population) firearm homicide rate of 3.8 per 100,000 people during 2009 (Kochanek, Xu, Murphy, Miniño, & Kung, 2011, Table 17). Typically, training to deal with potentially violent confrontations focuses on an officer's ability to select the appropriate response (e.g., shoot/don't shoot) and demonstrate appropriate tactics (e.g., tactical positioning, use of cover). However, little, if any, attention is paid to developing officers' ability to anticipate situational outcomes. Therefore, an additional goal of the current research is to gather information that can be used, in the future, to design training based on the anticipation skills, decision strategies, and processes of experienced officers (see Ward, Suss, & Basevitch, 2009).

To achieve the goals of this dissertation, two studies were conducted. The first study, consisting of three sub-studies, employed a temporal occlusion paradigm to identify video-based law enforcement scenarios that could be used to discriminate between experienced and less-experienced police officers, in terms of their ability to anticipate the outcome. This study also determined the optimal occlusion point for the discriminating trials. The second study—which employed the discriminating trials and occlusion points identified via the first study—investigated experienced-based differences in police officers’ ability to anticipate and assess situational outcomes, and also explored the relationship between these behaviors and response generation/selection.

Outline of Literature Review (Chapters 2 & 3)

In Chapters 2 and 3, the literature on the situation assessment and response phases of decision making are reviewed independently. To guide the reader, a more detailed description of each chapter follows.

The main focus of Chapter 2 is on research that has considered individual (and not team-based) situation assessment in representative and naturalistic settings. This is supplemented with research from less-dynamic domains (e.g., problem solving, chess) that has also considered relevant aspects of situation assessment. Specific attention is paid to the anticipation of situational outcomes. The review is divided into four subsections:

1. *Situation assessment in dynamic domains*. This subsection (a) introduces the major approaches that have addressed situation assessment in dynamic domains (e.g., recognition-primed decision making, Klein, 1989; situation awareness, Endsley, 1995a); (b) describes their conceptualizations of situation assessment; and (c) reviews

the associated evidence for skill-based differences in situation assessment.

Additionally, this subsection points out that situation assessment can be viewed as a process of hypothesis generation—a notion that is subsequently explored in greater detail.

2. *Hypothesis generation in ill-defined problems.* This subsection reviews research that demonstrates the utility of hypothesis generation tasks for understanding situation assessment. As relatively few empirical studies have examined hypothesis generation in naturalistic settings, this section focuses on studies that investigated hypothesis generation using text-based descriptions of ill-defined problems (e.g., automobile troubleshooting, Mehle, 1982). Aspects of hypothesis generation relevant to situation assessment in dynamic situations—such as the number of hypotheses generated and factors that stimulate the generation of new hypotheses—are highlighted.
3. *Perceptual anticipation.* This subsection (a) describes the concept of perceptual anticipation as future-oriented situation assessment in dynamic environments, (b) defines the scope of perceptual anticipation for the current study, and (c) reviews evidence for skilled perceptual anticipation in complex, time-constrained domains. The final part of this section considers theoretically-motivated investigations of skill-based differences in perceptual anticipation, one of which employed a hypothesis generation task to test predictions derived from the recognition-primed decision making model (Klein & Peio, 1989).

In Chapter 3, research is reviewed on the response generation/selection phase of decision making. This chapter is structured similarly to Chapter 2, which addressed situation assessment. The emphasis is on the processes involved in generating and

selecting, or deciding on, an appropriate course of action (rather than the extent to which an individual implements a chosen course of action effectively). The review consists of four subsections:

1. *Response generation/selection tasks to examine perceptual-cognitive expertise.* This subsection reviews studies that have employed response selection tasks to examine the underlying cognitive abilities that support superior performance. These studies typically focus on the quality of the selected response, rather than on the options that may have been generated prior to response selection.
2. *Response generation in ill-defined problems.* To complement the review of hypothesis generation in complex domains, this subsection reviews empirical research that has investigated response or “act” generation in ill-defined problems (e.g., Gettys, Pliske, Manning, & Casey, 1987). The review of these studies will highlight findings that inform a general understanding of processes involved in generating and selecting a course of action.
3. *Empirical studies of response-option generation in complex and dynamic domains.* This subsection reviews empirical research that has employed response option generation tasks to investigate skilled performance in complex and/or dynamic domains (e.g., Klein, Wolf, Militello, & Zsombok, 1995; Raab & Johnson, 2007). These studies typically suggest that decision makers generate “options” during response using a strategy that is consistent with the recognition-primed decision making model (Klein, 1989) and the Take-The-First heuristic (Johnson & Raab, 2003).
4. *Option generation/selection in the situation assessment and response phases of*

decision making. This subsection reviews studies that have examined, specifically, the relationship between the options generated during situation assessment and those generated during response. Although there are relatively few studies that have examined this relationship in dynamic situations, one of the studies reviewed in this section is particularly relevant, as it focused on situation assessment and response in law enforcement (Ward, Suss, Eccles, Williams, & Harris, 2011). Using verbal report data, Ward et al. found that during both phases of decision making, options were generated using a strategy consistent with long-term working memory theory (Ericsson & Kintsch, 1995). Furthermore, Ward et al. found that option generation in assessment and response interacted to produce a successful outcome.

CHAPTER 2

SITUATION ASSESSMENT IN DECISION MAKING

In this chapter, I review literature on situation assessment in dynamic domains, and then consider the use of hypothesis generation tasks as a means to understand situation assessment. The focus then turns to what has been termed the most interesting aspect of situation assessment—perceptual anticipation (see Sulistyawati et al., 2011; Suss & Ward, in press).

Situation Assessment in Dynamic Domains

In realistic situations, the decision making processes that result in the selection and implementation of a response are generally preceded by assessment of the situation (Wohl, 1981). In complex and dynamic domains, situation assessment is:

- A major feature of the recognition-primed decision making (RPD) model (Klein, 1989; Klein, Calderwood, & Clinton-Cirocco, 1986)—a descriptive model of expert decision making in naturalistic settings (e.g., firefighting, military command and control, neonatal intensive care). The RPD model defines situation assessment as “the sense of understanding what is going on during an incident” (Klein, 1989, p. 52), and describes how situation assessment is related to response selection.
- A process referred to in Endsley’s (1995a) model of situation awareness (SA) in dynamic decision making. SA, the primary construct of the model, is regarded as an individual’s state of knowledge about elements in the environment, and forms a major input to decision making. In the SA model, situation assessment is defined as “the process of achieving, acquiring, or maintaining SA” (Endsley, 1995a, p. 36); additionally, a theoretical model is developed to explain how this process works.

- Similar to the concept of sensemaking in naturalistic decision making. In the data/frame theory of sensemaking, Klein, Phillips, Rall, and Peluso (2007) defined sensemaking as “the deliberate effort to understand events . . . [that] is typically triggered by unexpected changes or other surprises that make us doubt our prior understanding” (p. 114). The theory explains the characteristics of sensemaking—such as its role in guiding response—and describes different types of sensemaking activities (e.g., building stories to account for data).
- Related to the use of situational assessment/probabilities in sport. Ward and Williams (2003) proposed a model to explain soccer players’ assessments of evolving patterns of play, in which players apply their experience to anticipate possible situational outcomes, and then assign probabilities to prioritize the anticipated outcomes in terms of their likelihood. According to the model, players’ then use these situational probabilities, together with other contextual information, to guide their response.

The following subsections describe situation assessment in the context of each model, and briefly review associated studies that provide evidence of skill in situation assessment.

Situation Assessment in Recognition-Primed Decision Making

The RPD model (Klein, 1989, 1993; see Mueller, 2009 for a computational instantiation) is a descriptive model of expert decision making, based largely on domain practitioners’ (e.g., fireground commanders, military commanders) accounts of critical decision incidents. According to the RPD model, domain practitioners first use their experience to recognize problem situations (e.g., a specific pattern/color of flames of a fire in an urban setting) as familiar or typical cases. Once recognized, the situation is

assessed by activating additional critical cues, plausible goals, expectations about how the situation might develop, and typical response actions. This process of recognition and situation assessment provides the operator with an understanding of the situation, and leads experienced decision makers to identify a single satisfactory course of action.

While readers may be most familiar with features of the RPD model that relate to the selection of a course of action, such as serial generation/deliberation of response options, Klein (1989) clearly distinguishes between generating/deliberating about possible states of the world during situation assessment and generating/deliberating about response options when formulating a course of action. In this section, the focus is on how decision makers understand the state of the world during situation assessment; response-option generation and selection will be addressed separately, in Chapter 3.

Regarding situation assessment, the RPD model posits that when environmental cues are relatively unambiguous, decision makers quickly develop an understanding or “take” on the situation (e.g., there is a fire in the laundry chute). Once initial classification (i.e., recognition) has occurred, situation assessment serves to develop an elaborated understanding of the specific instance by suggesting potential causes, identifying critical aspects to be considered, and highlighting potential future states and associated consequences. However, in more complex situations where the initial environmental cues are relatively ambiguous, or when new, inconsistent cues become available over time, decision makers may generate—either concurrently or serially—several hypotheses about the state of the world. In these relatively uncertain situations, each hypothesis represents a different assessment of the situation. Therefore, when several hypotheses are generated, decision makers must deliberate—again, either

concurrently or serially—between competing hypotheses in an attempt to settle on a single hypothesis that represents the most coherent assessment of the situation. RPD research has described specific examples of relatively complex situations in which decision makers generated, and then deliberated between several hypotheses during situation assessment (e.g., Calderwood, Crandall, & Klein, 1987; Klein, 1989).

RPD-related descriptions of situation assessment stemmed from research that employed retrospective interviews (e.g., critical decision method; Klein, Calderwood, & Macgregor, 1989) to examine critical incidents that occurred in naturalistic settings. Across a variety of domains, studies that employed the critical decision method have: (a) provided numerous examples of skilled situation assessment, (b) revealed critical cues used by domain practitioners to inform their interpretation of situations, and (c) identified some of the strategies employed by individuals engaged in situation assessment (e.g., Brezovic, Klein, & Thordsen, 1990; Calderwood et al., 1987; Crandall & Getchell-Reiter, 1993; Kaempf, Klein, Thordsen, & Wolf, 1996).

For example, Kaempf et al. (1996) interviewed naval officers who had made command-and-control decisions. Across 14 unique incidents, a total of 103 instances of situation assessment were identified to accommodate the changing dynamics of the situation. Two main situation assessment strategies were identified: feature-matching (87% of instances) and story building (12% of instances). Feature-matching was used to describe assessments of familiar situations based on cues present in the environment. When using this strategy, recognition occurred quickly and did not require further search for meaning. Story building, on the other hand, occurred when the available cues were insufficient for recognition. In these instances, the decision makers constructed plausible

stories (i.e., hypotheses) to connect and bring meaning to a series of cues or events. For example, an officer who observed enemy fighter jets circling his ship used information regarding their flight paths and radar activities to construct two stories: one where the jets were preparing to attack, and the second where the jets were only harassing his ship. Based on his assessment of the jets' activities, the officer concluded that it was more likely that they were engaged in harassing maneuvers, and were not preparing to attack the ship.

One limitation associated with retrospective interviews is that generally, decision makers report on different critical decision incidents. Therefore, this approach is not well-suited to examine skill-based differences in situation assessment because situation type may be confounded with expertise; expert practitioners may deal with different (e.g., more complex) situations than novices. To overcome this limitation, Calderwood, Crandall, and Baynes (1990) employed standardized building-fire scenarios to compare the situation assessments of urban firefighters with low ($M = 4.4$ years) and high ($M = 11.0$ years) levels of command experience. The scenarios were presented using a series of pictorial illustrations that included salient visual cues (e.g., type, size, construction, and age of the building) and depicted the development of the situation. Each illustration was supplemented with maps, narration, and radio communications. At 3–6 predetermined decision points per scenario, participants were asked to verbalize their thoughts out loud. Overall, experienced commanders generated proportionally more situation assessment statements and deliberated over alternate assessments more frequently than the less-experienced commanders. Moreover, the experienced commanders produced more situation assessment statements than action-related statements, whereas the less-

experienced commanders produced more action-related statements. The authors concluded that situation assessment plays a critical role in skilled decision making.

In summary, Klein and colleagues have provided a description of situation assessment and some evidence for skill-based differences in situation assessment. Although such research has highlighted different strategies used when assessing situations (i.e., feature matching, story building), the RPD model does not offer a more detailed explanation for how situation assessment occurs. It has, however, provided evidence that situation assessment is a continuous process that begins with detection of a problem, and can involve the generation and subsequent evaluation of alternate hypotheses for understanding the situation. The notion of situation assessment as hypothesis generation/evaluation has been investigated empirically to examine situation assessment in complex, but not strictly dynamic, domains (e.g., Gettys & Fisher, 1979; Klein & Peio, 1989). This related research will be considered in greater detail in subsequent sections.

Situation Assessment in the Model of Situation Awareness

According to Endsley (1995a), situation assessment is defined as “the process of achieving, acquiring, or maintaining situation awareness (SA)” (p. 36)—where SA is viewed as a state of knowledge or a product of situation assessment. More specifically, SA is defined as the “perception of the elements in the environment within a volume of time and space [Level 1 SA], the comprehension of their meaning [Level 2 SA], and the projection of their status in the near future [Level 3 SA]” (p. 36).

Research on SA has typically focused on quantifying this construct within operators. The most frequently used measure of SA is the Situation Awareness Global

Assessment Technique (SAGAT; Endsley, 1988; 1995b). Prior to administration, an a priori, goal-directed task analysis is conducted to identify task-critical situational information cues; a series of probe questions related to the three SA levels is then formulated to elicit participants' awareness of those cues (see Stanton, Salmon, Walker, Baber, & Jenkins, 2005). In a typical study, participants are engaged in simulated scenarios that are "frozen" either at random (e.g., Hogan, Pace, Hapgood, & Boone, 2006) or at critical points (e.g., Hogg, Follesø, Strand-Volden, & Torralba, 1995). At the freeze points, participants respond to several text-based SA probes. The simulation is then resumed and the procedure repeated until the scenario is complete. Response accuracy is measured by comparing the participant's responses to the state of the simulator at the freeze point, and/or by subject-matter expert evaluation. Variants of this method have been developed in which the probes are presented without pausing the simulator, which also permit response time to be recorded (e.g., Situation Present Assessment Method; Durso et al., 1998).

Although high levels of SA are considered necessary for skilled performance (Endsley, 2006), a search of the literature revealed few studies that examined skill-based differences in SA using actual domain practitioners. Moreover, researchers have often (a) reported a global rather than level-specific SA score (e.g., Hogan et al., 2006), (b) not clearly specified which levels are included in the global measure (e.g., Jones & Endsley, 2000), or (c) have not reported results for all three SA levels (e.g., Endsley, 1995b, studies 1 and 2; Strybel, Minakata, Nguyen, Pierce, & Vu, 2009). The following review of studies that have examined expert–novice differences in SA describes findings that relate to levels 1 and 2 SA (i.e., perception and comprehension); findings related to skill-

based differences in level 3 SA (i.e., projection) are addressed later, in the section on perceptual anticipation, along with other research that has focused on the future-oriented aspects of situation assessment.

Randel et al. (1996) used the SAGAT to assess the SA of 28 naval electronic warfare technicians. The technicians interacted with a high fidelity computer simulation, in which they observed, identified, and responded to radar signals emitted from potentially hostile sources. At two freeze points, technicians were asked to draw, from memory, the location and status (i.e., friendly, enemy) of emitters. After the scenario ended, the technicians answered questions about the threats they had observed, and their own ship's disposition (e.g., speed, prevailing weather conditions). Post-hoc, an objective measure of task performance was used to classify the technicians as novices ($n = 6$), intermediates ($n = 13$), or experts ($n = 9$). The experts ($M = 83\%$) were significantly better at recalling emitters than the intermediates ($M = 69\%$), who were, in turn, significantly better than the novices ($M = 46\%$). The experts also responded correctly to more questions about the threats they had encountered than the intermediates and the novices. Finally, the experts ($M = 4.8$ points) were significantly better at answering questions about their own ship's disposition than the intermediates ($M = 3.4$) and the novices ($M = 3.0$). The authors concluded that experts developed a better model of the situation than novices, and were better at assessing the situation from a tactical perspective.

Strater et al. (2001) measured the SA of army officers using the SAGAT and a subjective measure of SA—the Situation Awareness Behaviorally Anchored Rating Scale (SABARS). Seven skilled platoon leaders (i.e., captains) and seven novices (i.e.,

lieutenants with no platoon leader experience) completed computer-based military scenarios. Each officer performed the role of platoon leader, and deployed their forces to achieve prescribed mission objectives. SA was measured using the SAGAT at three predetermined points in two scenarios. At the freeze points, participants responded to SAGAT probe questions, including: (a) Which enemy locations are the strongest? (b) Which enemy element is your highest-level threat? and (c) Which friendly forces are currently exposed to enemy fire/attack? At the end of each scenario, a retired infantry officer—who had been playing the role of the commanding officer in the scenarios—rated the participant’s performance using the SABARS (20 items). An analysis of the participants’ SAGAT responses revealed that the skilled platoon leaders were significantly better than the novices at locating enemy troops and their own platoon, and at identifying the locations of the strongest enemy and the highest-level threat. The novices were significantly better at identifying the location of other friendly forces. A factor analysis of the rater’s SABARS data identified four factors that accounted for 67% of the variance in the 20 items. Skilled platoon leaders were rated as significantly better than novices on two of these factors: gathering information and following procedures, and focusing on the big picture. The authors concluded that the skilled platoon leaders had better comprehension of the threats posed by enemy forces, and suggested that training novices to focus on specific types of cues could lead to improved SA.

To explain how SA is produced, Endsley (1995a; see also Durso, Rawson, & Giroto, 2007) specified Kintsch’s (1988; van Dijk & Kintsch, 1983) construction-integration model as the mechanism that underlies situation assessment. In the construction-integration model, a constantly updated representation of the current

situation is formed—termed a situation model—by integrating situation-specific knowledge in long term memory with information currently available in the environment. An up-to-date situation model is a dynamic representation that includes the current state of the world in terms of the decision maker’s goals, and incorporates the temporal and dynamic aspects of the situation (e.g., how fast/slow the situation is developing, and what might happen in the future) (Endsley, 2000).

In summary, the SA model describes situation assessment as a comprehension-based process, and related research has focused largely on measuring the products of this process. Although the concept of SA is popular in human factors research (Durso & Sethumadhavan, 2008), only two studies were found that directly addressed the relationship between skill level and situational understanding in dynamic domains. Both studies assessed SA in adversarial tasks and found that skilled (i.e., better performing, more experienced) domain practitioners developed a more comprehensive understanding of the situation, especially with respect to their adversaries.

Although much of the research related to situation assessment in complex and dynamic domains stems from the RPD and SA research traditions, studies from other, complementary perspectives also address skilled situation assessment. The next section presents two additional examples—one from the sensemaking literature, and the other related to perceptual-cognitive skill in sport. These studies provide examples of situation assessment under more temporally constrained conditions.

Situation Assessment in Other Complex and Dynamic Domains

Sensemaking has been defined as “...a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their

trajectories and act effectively” (Klein, Moon, & Hoffman, 2006, p. 71). In a study that investigated skill-based differences in situation assessment from the sensemaking perspective, Zimmerman (2008) examined how police officers assessed a dynamic traffic-stop incident. Fifteen novice (i.e., 1-3 years of experience) and 15 experienced (i.e., 7-30 years of experience) police officers viewed video footage of an actual traffic-stop incident. The video depicted a dangerous, escalating situation in which the officer in the video questioned the occupants of a vehicle, and subsequently one of the occupants produced a gun. The event culminated with the officer shooting one of the occupants. The video was presented in four segments, and participants were instructed to observe the occupants’ behavior and the officer’s actions. After each segment, participants answered a series of questions about what they thought was happening on-screen, the cues they had attended to, and the appropriateness of the officer’s actions. Participants’ responses were coded using an inductive coding scheme. An analysis of the data revealed that the experienced officers referred to and interpreted more danger cues (e.g., the driver’s fidgeting hands), made more interpretations about the current situation, and made more predictions about impending events than the novice officers.

To explain how sensemaking occurs in surprising and anomalous events, Klein et al. (2007) proposed that data (i.e., cues) are interpreted using *frames*—explanatory structures that connect the data in plausible ways—and described several types of sensemaking activities. These activities, which are assumed to be conscious processes, include the comparison of multiple frames to determine the best match to the data, and changing frames when the existing frame is no longer adequate to explain the data. Note the similarity between these generative and evaluative activities and the concurrent and

serial evaluation of hypotheses described in the RPD model. Zimmerman (2008) presented evidence that participants used a variety of sensemaking activities to process information while observing the traffic-stop incident. For example, some participants initially observed that the vehicle's occupants were nervous, and hypothesized that the occupants were trying to hide a relatively minor crime (e.g., an unpaid traffic ticket). However, as the incident progressed and the occupants became even more nervous, participants generated an alternate hypothesis to explain the occupants' unusual behavior (e.g., the occupants were hiding something in the vehicle). Subsequently, when the occupants provided contradictory stories about drug paraphernalia that the officer found in the vehicle, participants typically abandoned their previous hypotheses and generated a different hypothesis to account for the available cues (e.g., the occupants were transporting a large quantity of illegal drugs). Once participants realized the seriousness of the situation, they generated predictions about what the occupants might do next (e.g., try to escape or try to harm the officer). This example demonstrates that in rapidly evolving incidents, situation assessment is an ongoing process that involves constructing an up-to-date understanding of the current situation based on cues available in the environment. Once a coherent understanding has been achieved, situation assessment can facilitate the prediction of impending events.

In fast-paced team invasion sports (e.g., soccer), players must assess the opposing team's developing pattern of play when considering their own defensive or attacking options. This type of situation assessment was originally referred to as the "use of situational probabilities" (Ward & Williams, 2003, p. 96). From this perspective, situation assessment involves the ability to prioritize relevant information, disregard

irrelevant information, and identify meaningful patterns of cues in the environment. Ward and Williams investigated how skill level and age affected situation assessment in a cross-sectional sample of soccer players ($n = 137$) aged between 9-17 years. Elite participants were players who trained for, and competed regularly in high-level national competition; sub-elite participants played only recreationally or at school level.

Approximately 14 elite and 14 sub-elite participants were recruited from each of five age groups: 9 years and younger (U-9), U-11, U-13, U-15, and U-17. Participants viewed short video clips of attacking soccer play, filmed from the perspective of the defending team. Each clip ended approximately 120 ms prior to the attacking player in possession of the ball passing to a teammate. At this point, participants highlighted the key attacking players who were in position to receive the ball, and ranked the highlighted attackers in terms of the threat they posed to the defense. Participants' responses were compared to those selected by a panel of expert coaches. Dependent variables were (a) the percentage of key players identified correctly, (b) the percentage of non-key players who were identified incorrectly as key players, and (c) a threat-ranking score that was determined by the number of correct matches to the coaches' rankings.

An analysis of the data revealed that skill groups used different strategies. As the age of participants increased from 13 to 18 years of age, elite players primarily reduced the number of non-key players they identified as threatening, having identified more key players than the sub-elite group from an early age (i.e. 9 years of age). In contrast, while the sub-elite players eventually caught up to the elite players by 13 years of age in terms of the number of key players highlighted, they did not make a meaningful reduction in the number of non-key players identified as age increased. Moreover, while both groups

increased their ability to prioritize the threats posed by key attacking players between the ages of 9 and 15, the elite group was consistently better at prioritizing threats at each age. From these findings, the authors concluded that elite participants were more accurate at assessing the situation than sub-elite participants, and that this difference was evident from an early age.

To explain elite players' superior situation assessment, Ward and Williams (2003) (see also McRobert, Ward, Eccles, & Williams, 2011; Ward, 2002) proposed that over time, skilled individuals develop extensive, detailed, and domain-specific knowledge representations (i.e., prototypical patterns of attacking soccer play) that are stored in long-term memory (see also McPherson, 1999). These authors proposed that, to facilitate access to these representations in time-constrained situations, skilled individuals develop long-term working memory skills (Ericsson & Kintsch, 1995) that allow for the efficient retrieval of knowledge from long-term memory. In parallel to Endsley's interpretation of SA, they proposed that in dynamic situations, skilled individuals then construct up-to-date situation models (Kintsch, 1988) by integrating contextual information from the environment (e.g., the positions and movements of attacking players) with their stored knowledge representations. The situation model, therefore, contains situation-specific, contextualized knowledge that highlights and prioritizes salient cues while omitting irrelevant cues, resulting in accurate situation assessment.

Thus far, the overview of situation assessment in dynamic domains has demonstrated that the notion of understanding complex situations has informed several approaches to the study of skilled performance. However, approaches have varied in their conceptualization of situation assessment, which has been described as a set of processes

which lead to understanding, and as the state of understanding itself. Additionally, situation assessment has been viewed both as something that is necessary only when a problem has been detected (e.g., RPD and sensemaking models), and as an ongoing monitoring process during routine situations that aids in the detection of problems or anomalies (e.g., situation awareness model). These differences are, in part, a product of perspective. The RPD model addresses both the situation assessment and response components of decision making, often in relation to specific critical or problem incidents. Therefore, apart from identifying critical cues used in situation assessment, the RPD approach has tried to determine how meaning is constructed in a given situation. SA, on the other hand, focuses primarily on the situation assessment component of decision making, often in relation to continuous tasks, such as flying, air traffic control, military command and control, and electronic warfare. The emphasis in the SA research has been on determining the extent to which an individual is situationally aware—or in other words, the quantification of SA. This measurement-based approach has been used primarily to evaluate display designs and to assess the efficacy of training programs; in comparison to these applications, it has only infrequently been used to examine skilled situation assessment.

The exemplar studies related to RPD, SA, sensemaking, and perceptual-cognitive skill in sport established that skill level and experience affect the quality of situation assessment in complex and dynamic domains—the more experienced practitioners developed more accurate and comprehensive interpretations of the situation. Recall that the RPD model (Klein, 1989) introduced the notion that when decision makers assess relatively complex and uncertain situations, they generate and evaluate hypotheses about

the state of the world. A number of researchers interested in understanding how people make sense of complex, but not necessarily dynamic, problems have employed hypothesis generation tasks as a way to understand the situation assessment process. The next section focuses on the utility of hypothesis generation as a tool for investigating situation assessment.

Hypothesis Generation in Ill-Defined Problems

In complex situations with high levels of information uncertainty (i.e., ambiguous cues), decision makers often need to engage in hypothesis generation and evaluation in order to settle on a coherent situation assessment (Wohl, 1981). Hypothesis generation tasks have been used as a tool for understanding how people assess ill-defined problems. This subsection reviews such research to demonstrate the utility of hypothesis generation tasks as a method for investigating situation assessment, and highlights findings that are relevant to situation assessment in dynamic situations, including (a) conditions that lead to new hypotheses being generated, (b) the number of hypotheses generated for a problem, and (c) evaluation of alternative hypotheses. This review will focus on studies in which participants had to generate their own hypotheses about ill-defined problems—which closely parallels the generation of hypotheses in naturalistic environments. Studies in which participants selected hypotheses from pre-specified lists (i.e., provided by researchers) are less representative of situation assessment in the natural ecology, and therefore are not considered here.

Gettys and Fisher (1979) observed that in ill-defined problem situations such as medical diagnosis, multiple hypotheses (e.g., diagnoses) were generated in response to the available information (e.g., medical test results), and that a set of hypotheses could

change in size and/or content when new information became available. This is analogous to situation assessment in dynamic domains, where relevant cues are likely to become available (or be perceived) in a sequential manner over time. Based on work by Shiffrin (1970) and Newell and Simon (1972), Gettys and Fisher assumed that hypotheses are generated via a recursive memory search that is directed and controlled by an executive process. They proposed that assessments of hypothesis plausibility acted as inputs to the executive process and determined whether hypothesis generation should continue, or be terminated. Consequently, they predicted that when the plausibility of the correct hypothesis being contained within a set of hypotheses is low, at least one new hypothesis should be generated that better explains the available information.

Forty-two undergraduate students were presented with lists of six words, one word at a time. For example, one list—beef, fish, aerospace industry, citrus fruit, tourists, and Cypress—contained products and industries associated with the state of Florida. In this problem, Florida was the *target* (i.e., to-be-generated) state but the first three words were also consistent with an alternate state—Texas. After being presented with the word “beef,” participants were asked to generate plausible target states (e.g., Texas, Missouri, Nebraska). Next, participants assigned likelihood ratings to each state in their set; 100 points were assigned to the state they believed was most likely to produce “beef”, and then relative likelihood ratings were assigned to the remaining states. Following this, participants were presented with the next word: “fish.” They then estimated the plausibility (scale: 0-100) that any of their initial set of states could be consistent with both words (i.e., “beef” and “fish”). After providing this plausibility estimate, participants updated their set of plausible target states for “beef” and “fish” by adding/dropping states

to/from their initial set. This procedure was then repeated for the remaining words in the list; each participant was presented with three lists. The main dependent variables were the plausibility estimates for each hypothesis set, and the likelihood ratings for each hypothesis (e.g., state) in the set. The plausibility data revealed that as new words from the list were introduced, rather than generating hypotheses at a constant rate for each new word, participants were three times more likely to generate new hypotheses when their plausibility estimates decreased, compared to when their plausibility estimates increased. An analysis of the likelihood ratings revealed that 90% of new hypotheses generated in response to the second through sixth list words were rated at least 50% as likely as the most likely hypothesis—evidence that participants were not simply guessing or padding their hypothesis sets with low-likelihood hypotheses. The authors suggested that these data were consistent with use of a heuristic, whereby new hypotheses are only added to the hypothesis set if they are strong alternatives to the existing best hypothesis. Although participants in this study were prompted (i.e., by the researchers) to evaluate their hypothesis sets, the findings suggest that in situations that evolve over time, newly perceived information is evaluated in relation to the existing situation assessment(s); an alternate hypothesis is then likely to be generated only if the existing situation assessment(s) cannot adequately explain all of the available information.

In a study that was mainly exploratory in nature, Mehle (1982) investigated skill-based differences in hypothesis generation in an ill-defined task. Based on common sense rather than a particular theory or empirical precedent, Mehle expected that skilled individuals would generate more hypotheses than less-skilled individuals. Six experts (i.e., experienced automobile mechanics) and six novices (i.e., undergraduate students)

were presented with text-based descriptions of five automobile problems. For example, one problem included the following information: “The car is American with an eight cylinder engine and an automatic transmission; it is two years old and due for a tune up. The problem is that the car refuses to start. The engine turns over and there is a gas smell” (Mehle, 1982, p. 93). Participants thought aloud while generating all plausible hypotheses (i.e., causes) for each problem, and then indicated the probability (scale: 0–100) that their hypothesis set contained the true (i.e., actual) hypothesis. Overall, experts and novices did not differ significantly in the number of hypotheses generated for each problem ($M_{\text{experts}} = 3.36$, $M_{\text{novices}} = 3.43$) or in their assessments of plausibility ($M_{\text{experts}} = 0.675$, $M_{\text{novices}} = 0.692$). On average, each participant generated only 19% of the entire (i.e., pooled) list of hypotheses generated by all participants. The only skill-based difference noted was that the experts generated hypotheses that were more specific than those generated by the novices. Note, however, that the quality (i.e., accuracy) of hypotheses was not addressed in this study. The findings suggest that in ill-defined situations, rather than generating a large set comprising all possible hypotheses, individuals interpret the available information in a way that constrains the number of hypotheses generated. Mehle suggested that generating only a small number of hypotheses might be an adaptation designed to cope with limited working memory capacity (see also Thomas, Dougherty, Sprenger, & Harbison, 2008).

In ambiguous situations, such as those that can occur in medical diagnosis, the number and type of hypotheses generated may also be influenced by the base rates of the plausible, underlying disorders. Weber, Böckenholt, Hilton, and Wallace (1993) predicted that hypotheses consistent with higher base-rate disorders would be generated

more frequently, and would be rated as more likely, than hypotheses consistent with lower base-rate disorders. Eighty-four doctors read brief, written descriptions of three general-practice cases that contained cues related to one high and one low base-rate diagnosis; in all cases, the high base-rate diagnosis was associated with less severe clinical consequences than the low base-rate diagnosis. After reading the description, the doctors generated plausible hypotheses (i.e., diagnoses), which they then ranked in order of likelihood. On average, the doctors generated four hypotheses per case. Hypotheses consistent with the high base-rate diagnosis were generated more frequently than the low base-rate diagnosis (96% versus 70% of trials, respectively). Hypotheses consistent with the high base-rate diagnosis were also ranked as the most likely more frequently than those consistent with the low base-rate diagnosis (74% versus 8%, respectively).

The authors concluded that while the doctors were sensitive to base rate information (i.e., listed hypotheses consistent with the high-base rate diagnosis first), they typically generated at least one low base-rate (but high-severity) hypothesis in their set. The authors suggested that this strategy—of considering at least one severe outcome—could benefit patients by reducing doctors' chances of missing a relatively harmful disorder. A similar strategy may also be adopted by decision makers in other domains. For example, a police officer dealing with a suspect who presents a low level of threat (e.g., complies with the officer's commands) may nonetheless consider the possibility that the suspect could become violent—even if the officer believes there is a low likelihood of this occurring. Doing so may serve to prime an officer's response, should the suspect actually turn violent. Therefore, when asking police officers to generate hypotheses about dynamic law enforcement situations, eliciting likelihood and threat

ratings for each hypothesis may be useful in discriminating between different types of hypotheses (e.g., low likelihood/high threat versus high likelihood/low threat).

To summarize, the studies that examined hypothesis generation using text-based descriptions of ill-defined problems demonstrated the utility of hypothesis generation as a tool for examining situation assessment, and informed our understanding of situation assessment in several ways. First, the studies found that individuals generated relatively few hypotheses—often less than five—when assessing problem situations. This suggests that when assessing a given situation, rather than generating all possible alternative explanations, individuals generate only those hypotheses they believe to be most relevant (or have the most utility). Second, there is evidence that as more information becomes available, individuals update their hypothesis sets by generating new hypotheses that are consistent with the totality of the information, and/or dropping unlikely or inconsistent hypotheses from their current set. In some situations, however, individuals may benefit from retaining at least one low-likelihood (but high-severity) hypothesis. Third, the studies by Gettys and Fisher (1979) and Weber et al. (1993) suggest that likelihood is at least one dimension by which hypotheses can be ranked, and that likelihood ratings may be useful in understanding an individual's hypothesis generation strategy. The next section focuses on what is, arguably, the most interesting aspect of situation awareness—the ability to predict situational outcomes.

Perceptual Anticipation

This section addresses the notion of prediction in dynamic domains, and identifies key aspects that will inform the study of skilled performance in law enforcement. It begins by providing a brief overview of some constructs related to prediction, and in

particular perceptual anticipation, that have been used to assess skilled performance. As different terms have been used to describe related phenomena, care is taken to define the scope of the term “perceptual anticipation.” Next, evidence of skilled perceptual anticipation in three dynamic domains is presented. Rather than provide an exhaustive review, the aim here is to demonstrate the relevance of perceptual anticipation for skilled perceptual-motor performance, and to emphasize the methods that have been used to measure it.

Perceptual Anticipation: Related Constructs

When the concept of situation awareness (SA) first emerged in human factors, it was used in the context of a “dog fight” between aviation fighter pilots, where one aviator tried to “get inside the head of the other”—to predict what their opponent might do next—as a means to gain a tactical advantage (R. Hoffman, personal communication). Although my intention is not to address the debate over definitions of SA, this early definition placed the emphasis close to the position that skilled prediction provides a good opportunity for investigating expertise. Using simple laboratory tasks to investigate the role of prediction in skilled action, Poulton (1957) coined the term *perceptual anticipation* to refer to the ability to predict the future state of a situation in order to coordinate and adjust an appropriate response relative to that future state. Such judgments are made, for instance, when the future trajectory or position of a target is not prescribed but could be determined from experience or based on statistical properties of the environment.

Poulton’s (1957) definition of perceptual anticipation was primarily concerned with situations in which an in-event, adaptive response was required based on changes in

the immediate environment (e.g., tracking a target to a predictable location). In this dissertation, this definition is extended to include predictions made *in situ* while adapting to dynamic changes in complex environmental events, including when those adaptations actively change the environment. Henceforth, the term *perceptual anticipation* is used as a unifying concept to refer to the ability to generate real-time, mental projections about how a situation may unfold in the immediate or very near future (e.g., seconds). Although some researchers have reserved the term *mental projection* to refer, primarily, to a process in which individuals engage to “figure out” how events could unfold or if a course of action might work (e.g., Klein, 1989), one of the main arguments of this dissertation is that mental projection can occur “on-the-fly”—in the form of perceptual anticipation of situational events—and does not necessarily require individuals to take a “time out” to do it (see Kintsch, 1988; Poulton, 1957).

Although most expertise researchers have focused on aspects of expertise other than perceptual anticipation (e.g., memory), the literature is replete with examples of skilled performers employing this skill to great effect. For instance, expert pilots have been shown to “fly ahead of the plane” (e.g., Doane, Sohn, & Jodlowski, 2004), advanced drivers anticipate hazards in the road ahead (e.g., McKenna & Horswill, 1999), athletes are highly skilled at “reading the game” (see Williams & Ward, 2007), and police officers can anticipate perpetrators’ actions before they occur (Ward et al., 2011). Data from this research suggest that superior performers routinely think ahead— in a mode that might be called direct apprehension (in contrast with deliberative reasoning)—about what is going to happen next. Importantly, they make use of this information to guide their behavior.

A number of phenomena related to perceptual anticipation have been studied in complex domains under various guises, such as anticipation, anticipatory thinking, foresight, hazard perception, prediction, and projection. Each construct has stressed, to varying degrees, the importance of perceptual anticipation as a precursor to generating courses of action and/or as part of the perception-action cycle (Neisser, 1976; Smith & Hancock, 1995). The following subsections review key methods that have been used to investigate perceptual anticipation, highlighting how these approaches have informed our understanding of the role of perceptual anticipation in skilled performance.

Perceptual Anticipation in Complex Domains

While anticipating an opponent's next move has been an integral part of military stratagem since Sun Tzu's *Art of War* (Tzu, trans. 1983), in modern times, de Groot (1965) was among the first to study the perceptual-cognitive basis of predicting an opponent's intentions using scientific methods. In his seminal research, de Groot used a move-selection paradigm as a means to study expert thinking in chess, where expert chess players were asked to select their next best move based on the current configuration of chess pieces. Compared to less-skilled players, experts chose better moves for themselves. Although players did not differ in the depth of their search, experts considered higher quality moves in more detail through a process of progressive deepening. While subsequent research focused on the debate between recognition and search as a mechanism supporting personal move selection (Chabris & Hearst, 2003; Gobet & Simon, 1996b), de Groot's research also highlighted that experts progressively deepened their search for better moves in *plies* (e.g., potential moves for oneself and predicted countermoves of their opponent).

Quite apart from chess, perceptual anticipation has been investigated as part of several distinct research traditions, each approaching the construct from different, yet equally informative perspectives. Next, I review research in three main areas: anticipation in sport, prediction-specific components of situation awareness in dynamic domains (e.g., air traffic control), and hazard perception in driving. Studies in these domains have examined perceptual anticipation in complex tasks, where individuals work to achieve goals within a limited time window. In light of the dearth of studies related specifically to law enforcement, studies in these areas will provide evidence for the relationship between perceptual anticipation and skilled performance, and highlight the methods used to examine perceptual anticipation in dynamic and complex environments.

Perceptual anticipation in sport. The study of anticipation skill in a dynamic context has a rich history in the sports-expertise and -training literature (e.g., Abernethy, 1987; Haskins, 1965; for a review see Williams & Ward, 2007). In many sports, performers have a very limited time in which to initiate a response. For example, in 1-player versus 1-player situations in fast ball sports (e.g., tennis serve, soccer penalty kick, baseball pitch), to ensure a successful interception of the ball, players are often forced to anticipate the outcome of their opponent's action by interpreting the probabilistic postural cues of the opposing player that are available prior to the ball leaving their opponent's racket/bat/limb. Initiating a response after this "contact" point and when other more deterministic and confirmatory cues are also available (i.e., ball flight trajectory), on average, results in insufficient time to react and intervene successfully (see Williams, Davids, & Williams, 1999).

In individual and team sport settings, researchers have often used a temporal occlusion paradigm to investigate skilled anticipation. For example, Abernethy (1990) investigated the extent to which skill level affected squash players' ability to perceptually anticipate the direction and force of an opponent's squash strokes. Ten skilled (i.e., state-ranked) and 15 novice (i.e., non-competitive) squash players viewed video simulations, taken from a first-person perspective, of an opposing player hitting strokes from different positions on the court. The video stimuli were occluded at one of five occlusion points: 160 or 80ms prior to racket/ball contact, at racket/ball contact, 80 ms after racket/ball contact, or after the stroke was completed. Participants observed the stimuli and, at the point of occlusion, were required to verbalize both the direction (i.e., down the line/cross court) and force (i.e., long/short) of the stroke. Overall, the skilled players were significantly more accurate in anticipating the direction and force of strokes, compared to the less-skilled players. Further analysis of the data revealed that skilled players' used both early (i.e., between 160 ms and 80 ms prior to racket/ball contact) and late (i.e., between 80 ms after contact and stroke completion) cues to anticipate shot direction, while less-skilled players were only able to make use of the late cues. These findings suggest that the skilled players were attuned to important cues in the environment, which they used to perceptually anticipate their opponent's actions.

Perceptual anticipation in sport has also been investigated using a reaction time paradigm. For example, Williams, Ward, Knowles, and Smeeton (2002) investigated the extent to which skill level affected tennis players' ability to perceptually anticipate tennis ground strokes. Eight skilled tennis players with tournament experience and 8 recreational tennis players (i.e., with no tournament experience) viewed video

simulations, taken from a first-person perspective, of an opposing tennis player hitting a groundstroke to one of four locations on the court. Participants held a tennis racket and began each trial standing on two pressure-sensitive pads. They then viewed the video stimuli on a projection screen and simulated their return shot by stepping to intercept the shot and swinging their racket. The video was occluded when participants lifted a foot from one of the pressure-sensitive pads, and reaction time was measured from the start of the clip until the point of occlusion. Anticipation accuracy was determined by whether the direction of participants' movement (i.e., laterally left/right, directly forwards/backwards) corresponded with the actual destination of the opponent's simulated groundstroke. Analysis of the data revealed that while skilled and less-skilled players did not differ significantly in the accuracy of their anticipations (68.4% versus 64.5% correct, respectively), the skilled players reacted significantly faster (by approximately 140 ms) than the less-skilled players. Analysis of eye movement data collected during the trials revealed that skilled players fixated longer on central cues related to the trunk/hip and head/shoulder regions, while the less-skilled players fixated longer on more proximal cues, including the racket and the ball. This suggests that skilled players have developed visual search strategies that enable them to focus on briefly-available postural cues that are displayed earlier in their opponent's movement.

In team game settings, where multiple players interact to produce a coordinated outcome, patterns of cues generated by the coordinated movements of multiple players are also used to anticipate the outcome of the play. For instance, Williams and Davids (1998) created video simulations of 3 versus 3-player soccer plays, filmed from the perspective of a fourth, rear defender. Each simulation ended with one of the three

attacking players either passing to another attacker on his left or right side, dribbling straight toward the rear defender, or shooting at goal (i.e., over the defender's head).

First, using a reaction time paradigm similar to that employed by Williams et al. (2002) for tennis, skilled and less-skilled soccer players anticipated the direction of the attacker's play. As in the tennis study, while the two groups did not differ significantly in their anticipation accuracy, the skilled soccer players initiated their movement significantly earlier than the less-skilled players. However, it was not clear whether the skilled players were relying more on the postural cues of the player with the ball, or on the actions and movements of the other players.

To examine which cues were more important for anticipation in multi-player situations, Williams and Davids (1998) conducted a subsequent experiment using a spatial occlusion paradigm. Two versions of the 3 player versus 3 player simulations were created by editing the original videos. In one version, a black mask was used to occlude everything except for the player with the ball. In the other version, all six players were visible. Skilled and less-skilled players viewed masked and unmasked stimuli and anticipated the direction of the attacker's play. Analysis of the reaction time data revealed that skilled players reacted faster than the less-skilled players in both occlusion conditions. However, skilled players' anticipation accuracy was significantly degraded when information about the other players was occluded, while less-skilled players' accuracy did not differ significantly across the occlusion conditions. These data suggest that skilled players are able to anticipate the intentions of the player with the ball, based not only on the postural cues of that player alone but also on the actions and movements of other players who form an integral part of the emerging pattern of play. When critical

cues are removed from the display, expert anticipation is degraded toward more novice-like performance (for a review see Williams & Ward, 2007).

Collectively, the research in sport suggests that mastering the skill of perceptually anticipating future events is key to attaining a skilled level of performance in this domain. Sports situations that involve interaction with an opponent(s) under constantly evolving circumstances are analogous to a range of other societally-relevant situations in which highly developed perceptual anticipation is linked to skilled performance. I now review how perceptual anticipation has been addressed by researchers in these complementary, dynamic domains, under the general guise of situation awareness.

Projection in dynamic domains. Recall that Endsley's (1995a) model of situation awareness (SA) described the ability of a human actor to anticipate a near future state—projection (or level 3 SA)—as the most advanced level of SA. Projection is assumed to be built on the lower levels of SA (i.e., perception and comprehension) and to precede the decision-making and action components of performance (Endsley, 1995a). While the emphasis on projection in SA research has been immensely variable, SA probes frequently include questions that allow an assessment of perceptual anticipation skill. For instance, Endsley, Sollenberger, Nakata, and Stein (1999) presented air traffic controllers with computerized simulations, and using the SAGAT, asked them: (a) Which pairs of aircraft will lose separation if they stay on their current (intended) course? (b) Which aircraft must be handed off to another sector/facility within the next two minutes? (c) Which aircraft will violate special airspace separation standards if they stay on their current (intended) paths? and (d) Which aircraft will weather be an impact on in the next five minutes along their current course?

However, while much research exists, for instance, that compares methods of data collection or the effect of display design on SA (e.g., Endsley, 1995b; Endsley et al., 1999), rather surprisingly, there is a limited amount of research that specifically describes the skill-based differences in mental projection and/or the predictive ability of experienced/skilled operators. Consequently, despite the use of innovative methods (e.g., SAGAT), only a small number of SA-related studies have examined skilled operators' ability to anticipate future states of the system.

In a study on air traffic control, Durso et al. (1998) examined whether different measures of SA and workload could be used as predictors of performance. Experienced controllers completed five scenarios; in two of these (one SAGAT, one Situation Present Assessment Method) they answered questions about the present (e.g., Which aircraft has the lower altitude, aircraft A or aircraft B?) and future (e.g., Will aircraft X and aircraft Y be traffic for each other?) state of the system. Workload, and past and present SAGAT queries combined to account for 74% of the variance in performance—measured as the controller's efficiency in clearing the airspace. However, their analyses also revealed that superior performance was positively associated with the speed and number of correct responses to future queries, but negatively related to the number of correct responses to present queries. These data suggest that focusing on the future is an important component of superior air traffic control performance (but see Durso, Bleckley, & Dattel, 2006).

In a study that focused on the importance of the projection component of SA, and reported these results independently of the global score, Sulistyawati et al. (2011) examined experienced fighter pilots' ability to predict what would happen next in a PC-based air-combat simulation. Participants were assessed using the SAGAT. The

simulation was frozen either 10 seconds after enemy aircraft appeared on radar or when the pilot was within 25 nautical miles of enemy aircraft. Immediately after occlusion, participants responded to a number of SA probes including two that specifically assessed projection: (a) Will you be in a position to take shots at an enemy aircraft in the next 10 seconds? and (b) Will the enemy aircraft take shots at you within the next 30 seconds? The data indicated that participants correctly predicted their own and their enemy's ability to get into a position to shoot 67% of the time, however, the least experienced pilots were unable to respond appropriately to either comprehension (i.e., Level 2) or projection (i.e., Level 3) assessments. A regression analysis that included each level of SA and a measure of overconfidence bias revealed that the ability to project into the future, together with less overconfidence, predicted mission performance—the number of times the pilots were actually shot at by the enemy aircraft.

Assessments of SA, including projection, have also been conducted using subjective self-report and observer rating scales. In one such study, Matthews and Beal (2002) investigated cadet platoon and squad leaders' SA during a military field training exercise. At the conclusion of each exercise, the platoon and squad leaders self-reported on their level of SA using the Mission Awareness Rating Scale (Matthews, Beal, & Pleban, 2002), and one or two experienced officers or non-commissioned officers rated participants' observable behaviors using the Situation Awareness Behavioral Rating Scale (SABARS; Strater et al., 2001). Both scales included a rating of the participants' ability to predict future outcomes. Although the platoon leaders rated themselves as significantly better at predicting the outcome of the situation compared to the squad

leaders, the experienced observers did not rate platoon and squad leaders differently. More objective measures of prediction were not recorded.

Collectively, SA methods are well-established and, when level 3 probes are used and reported separately, are directly applicable to the measurement of perceptual anticipation. SA research that has focused specifically on projection in dynamic domains indicates that the ability to perceptually anticipate the future status of elements within the environment is associated with higher levels of performance, although SA research has not delved deeper into phenomena of mental projection on-the-fly.

Perceptual anticipation has also been examined in automobile driving, where it is commonly referred to as *hazard perception*. Although hazard perception research developed quite separately from the research on anticipation in sport and SA in dynamic domains, these approaches offer complementary insights into skilled anticipation. Next, I briefly review the hazard perception literature, highlighting the methods used to investigate perceptual anticipation in driving and noting relevant findings.

Hazard perception in driving. The ability to perceptually anticipate the immediate future state of the world has been studied in the domain of automobile driving as a means to understand hazard perception. Hazard perception has generally been defined as the ability to “read the road” ahead and to anticipate and identify potentially dangerous traffic situations (McKenna & Crick, 1994). It is a useful predictor of accident involvement: less-skilled drivers are much more likely to detect hazards late because they do not anticipate the road ahead (Horswill & McKenna, 1999; McKenna & Horswill, 1999; see also Rumar, 1990).

A response time paradigm was originally developed to measure hazard perception in driving. For instance, McKenna and Crick (1994) developed a hazard perception test in which they presented participants with a 30-minute continuous video, comprised of sequences of road traffic filmed from a first-person perspective within a car. The video contained 35 commonly occurring road hazards to which drivers would have to respond, as well as periods where no hazards appeared. While viewing the video, participants pressed a response button each time they perceived something that might turn in to a dangerous situation. Dangerous situations were defined as those where the participant thought an accident or near accident might happen next, and where they considered it necessary to take some kind of evasive action (see McKenna, Horswill, & Alexander, 2006). The results showed that highly experienced police driving instructors (i.e., with 11 or more years' driving experience, and who were qualified as Class One Police drivers) anticipated the hazards significantly earlier than less-experienced civilian drivers (i.e., with 11 or more years' driving experience) who, in turn, responded earlier than inexperienced drivers (i.e., with up to 3 years' driving experience). Additionally, the experienced police driving instructors identified significantly more of the hazardous incidents than the inexperienced drivers. These results have been replicated several times since the original study (see McKenna & Horswill, 1999; McKenna et al., 2006).

McGowan and Banbury (2004) adapted McKenna and Crick's (1994) hazard perception test by incorporating the SAGAT to examine the relationship between SA and anticipation. Moreover, they examined whether the process of interrupting participants to provide SA queries (that included projection probes) affected their ability to anticipate the outcome. Throughout the hazard perception video test, which contained 70 hazardous

events, participants in an SA query condition were interrupted 32 times when the video was paused for 10 seconds while they responded to a multiple-choice SA probe. The number of correct responses to SA queries was positively and significantly correlated with the ability to anticipate the hazards. Participants in the SA query condition anticipated potential hazards on both interrupted and non-interrupted trials approximately 2 seconds earlier than participants in a control group that anticipated all events without interruption. However, on the trials where the SA query group were interrupted, participants took slightly (i.e., approx. 0.5 seconds) longer to anticipate hazards suggesting that there was a general carry-over effect that benefited anticipation on the non-interrupted trials. Although, these authors did not report level 3 SA results or examine the relationship between projection and anticipatory performance per se, these data provide some preliminary indication that the two measures were convergent in identifying anticipation skill.

Consistent with research on anticipation in other domains, the hazard perception research that has assessed skill-based differences in perceptual anticipation indicates that skilled drivers are able to predict that dangerous situations are likely to ensue, and are better than novices at reading the road ahead (see also McKenna and Horswill, 1999; cf. Jackson, Chapman, & Crundall, 2009; Vogel, Kircher, Alm, & Nilsson, 2003). Moreover, the available data suggest that when relatively inexperienced drivers are primed to attend to aspects of the situation that may develop into hazards, they are better able to anticipate dangerous situations (McGowan & Banbury, 2004).

Overall, the research on anticipation in sport, situation awareness in dynamic domains, and hazard perception in driving suggests that perceptual anticipation is an

important component of superior performance in complex domains that, similar to law enforcement, require decisions to be made in constantly changing and time pressured situations. However, while this research described expert–novice differences in perceptual anticipation, reasons for these differences were not elucidated. To the best of my knowledge, there are few studies that have tested theoretical explanations for skill-based differences in perceptual anticipation. The next subsection reviews those studies: Two of those studies were conducted in a relatively static domain (chess; Klein & Peio, 1989; Jastrzemski, Charness, & Vasyukova, 2006), and one study addressed anticipation in a dynamic domain (soccer; Ward, Ericsson, & Williams, 2012). The Klein and Peio, and the Ward et al., studies are especially pertinent in the context of the current proposal, as they tested predictions derived from the RPD and LTWM models, respectively. Additionally, those studies employed an option-generation paradigm similar to the hypothesis generation tasks used in the studies of ill-defined problems.

Theoretically-Motivated Studies of Skill in Perceptual Anticipation

Recall that one of the main messages from the RPD model (Klein, 1989) was that experts' superior ability to recognize situations allows them to generate a satisfactory, if not a good, course of action as the first one considered. However, in addition to explaining how individuals make decisions about what course of action they might take themselves, Klein and Peio (1989) suggested that the RPD model would be equally relevant during situation assessment, when individuals had to predict (or perceptually anticipate) what an opponent might do next. Klein and Peio modified de Groot's (1965) original move-selection paradigm to create a prediction-based, situational option-generation paradigm in which chess players were asked to predict the next moves of

other, expert players. Based on the RPD model, Klein and Peio expected that, compared with novice players, proficient players would: (a) generate fewer prediction options per trial, (b) select the correct move (i.e., actual move played) as their final prediction more frequently, (c) include the correct move in their set of predictions more frequently, and (d) generate the correct move earlier in their set of predictions. Twenty-four novice (i.e., ELO rating ≤ 1300) and 10 proficient chess players (i.e., ELO rating ≥ 1700) were shown a chessboard configuration from an actual expert-level game, with the pieces arranged 10 moves in. Next, they identified the plausible moves that the opposing expert player might make next, and then highlighted the one they thought the expert would actually make. Participants then saw the actual move made by the expert and the task was repeated for the 20 subsequent moves (i.e., 10 moves for each expert player). Consistent with the RPD model, proficient players generated fewer options per trial ($M_{\text{proficient}} = 2.2$, $M_{\text{novice}} = 2.5$), and predicted the experts' actual move significantly more often than the novice players did ($M_{\text{proficient}} = 38\%$ of trials, $M_{\text{novice}} = 23\%$). The proficient players also generated the correct move in their set of predictions more frequently than the novice players did. Finally, when the proficient players correctly predicted the experts' next move, they generated the correct move as their first prediction option a significantly greater proportion of the time ($M_{\text{proficient}} = 59.5\%$ of trials, $M_{\text{novice}} = 41.3\%$). The authors concluded that the major features of the RPD model (i.e., few options are generated, and the first option generated is typically a good option) described option generation during situation assessment, and that differences in option generation/selection are associated with perceptual anticipation skill. This study also demonstrated the utility of a prediction-based, situational option-generation paradigm for revealing skill-based differences in the

generation and selection of options during perceptual anticipation.

Jastrzemski et al. (2006) assessed the effects of expertise and age on anticipation in chess; my review focuses on the most relevant aspect of that study: the theoretical basis for the relationship between expertise and anticipation. According to Gobet and Simon's (1996a) theory of chunks and templates, expert chess players have access to an extensive store of chess patterns. Those patterns are thought to include information regarding the threat relationships between chess pieces (e.g., how susceptible a king is to being placed in "check"). Jastrzemski et al. hypothesized, based on LTWM, that expert chess players employ cues—held in working memory—to rapidly retrieve those encoded relationships, thus allowing them to anticipate threats faster than novice players. Furthermore, they hypothesized that experts would be able to assess threats-of-check as quickly as the actual presence of a check. To assess skill-based differences in anticipation performance, 20 expert (i.e., ELO ratings of 2,195–2,540), 19 intermediate (i.e., ELO ratings of 1,700–2,060), and 20 novice (i.e., with no ELO rating) chess players performed two detection tasks. In both tasks, players viewed patterns of chess pieces (i.e., subsets of chunks) that were displayed on a computer screen. Each pattern was represented on a 4 square \times 4 square section of a chessboard, and consisted of two chess pieces—one of which was a king. On *check* detection trials, participants determined whether or not the king was in check—a relatively simple task in which all of the information required to make a decision was contained within pieces displayed. On *threat-of-check* detection trials, participants determined whether, in a single move, the king could be placed in check; as the pieces represented only a subset of a chunk, participants needed to mentally simulate other parts of the chunk in order to make a threat-of-check determination. For

both detection tasks, accuracy and response time were recorded; accuracy was high (i.e., > 98%) for all groups. An analysis of response times for the accurate trials revealed a significant Skill \times Task interaction: expert players' response times did not differ significantly across the two tasks, while the novice's took significantly longer to respond on threat-of-check trials compared to check trials. Additionally, expert players responded significantly faster—on both tasks—than did the novices. Jastrzembski et al. concluded that the results supported the hypothesis that experts could rapidly access stored templates to anticipate possible outcomes, and that their perceptual advantage—evident from the speed with which they assessed patterns—was an important component of skilled chess play.

Ward et al. (2012) built on Klein and Peio's (1989) work by examining option generation during perceptual anticipation in a dynamic domain: soccer. Ward et al. contrasted hypotheses drawn from the RPD model and Take-The-First heuristic (TTF; Johnson & Raab, 2003) with those derived from LTWM theory. TTF, which is discussed in more detail in Chapter 3, offered an explanation for why generating few options and selecting the first option (i.e., behaving in accordance with the RPD model) should result in better performance: it proposes that there is a negative correlation between the *total* number of options generated, and the quality of the selected option. LTWM theory, on the other hand, posits that superior anticipation is the product of a rich, accurate, and up-to-date situational model. Therefore, based on LTWM theory, Ward et al. expected that when multiple task-relevant options exist in the environment, skilled performers would generate more of those options, and fewer task-irrelevant options, than would novices. Also based on LTWM theory, and in contrast to TTF, Ward et al. expected that the

number of task-relevant and task-irrelevant options would be positively and negatively correlated, respectively, with the quality of the selected option. Finally, it was expected that skilled participants would be better at prioritizing task-relevant options in terms of threat.

In Ward et al.'s (2012) second and third experiments, skilled and less-skilled soccer players viewed, from the perspective of a defensive player, short, temporally-occluded clips of attacking soccer plays. The clips were occluded approximately 120 ms prior to the attacking player who possessed the ball either (a) kicking at goal, (b) passing to another attacker, or (c) running while maintaining possession of the ball. At the point of occlusion, participants first anticipated what the attacker would do next (i.e., shoot at goal, pass, or run) and then generated situation assessment options to represent what the attacker *could* (cf. would) do next. After generating assessment options, participants ranked those options with respect to the threat posed to the defense. Participants completed 18 trials. Correct anticipations were those that matched what the attacker actually did after the occlusion point (i.e., if the clip had not be occluded). Expert soccer coaches determined the task-relevant options, and optimal threat-ranking of those options, for each clip. A threat prioritization score was calculated by comparing participants' rankings to the experts' criterion rankings.

As expected, skilled players ($M = 15.13$) anticipated the actual outcome on significantly more trials than did the less-skilled players ($M = 10.25$). Consistent with LTWM theory, skilled players generated significantly more task-relevant, and fewer task-irrelevant, options than did the less-skilled players. Skilled players also prioritized their task-relevant options significantly better than did less-skilled players. Also consistent

with LTWM theory, the number of task-relevant and task-irrelevant options were positively and negatively correlated, respectively, with anticipation performance. Ward et al. (2012) concluded that skilled anticipation performance was supported by LTWM-type mechanisms, as opposed to the mechanism proposed by TTF.

Summary

In summary, the review of situation assessment in complex and dynamic domains provides support for using the ability to anticipate situational outcomes as a principled basis for understanding professional expertise in law enforcement by:

- Describing the role of situation assessment in complex and dynamic domains, and presenting evidence of skill in situation assessment.
- Highlighting the use of option (e.g., hypothesis) generation tasks as a method for investigating situation assessment.
- Emphasizing the importance of perceptual anticipation, as the future-oriented aspect of situation assessment, in constantly evolving situations, and presenting evidence of skill in anticipating future situational outcomes in sport, dynamic domains, and driving.
- Presenting evidence that skill in predicting situational outcomes in a complex task can be understood through the use of a prediction-based, situational option generation paradigm.

Having reviewed literature related to situation assessment, the following section provides a complementary review of the response generation/selection phase of decision making.

CHAPTER 3

RESPONSE PHASE OF DECISION MAKING

One of the hallmarks of expertise is that skilled performers make better decisions about how to respond to situations than less-skilled performers. In the judgment and decision-making literature, lab-based studies of decision making have typically employed well-defined problems and focused on individuals' ability to select one option from a list of specified alternatives. However, in complex, uncertain, and dynamic situations such as those encountered by law enforcement officers, domain practitioners must first generate at least one response option for themselves prior to selecting a course of action. Formal, decision analytic models such as subjective expected utility and multi-attribute utility theory specify methods for selecting one option from among alternatives, but are silent on the issue of option generation (Yates, 2001). Other decision theories that do address the issue of option generation, such as decision analysis (e.g., Howard, 1968), suggest that generating an exhaustive or near-exhaustive set of alternatives is a prerequisite for making a high-quality decision. Although decision analysis has been applied with success in a variety of domains (e.g., management science, computer science), this method is demanding in terms of both the time taken and computational power needed to complete such analyses—two resources that are in short supply when police officers are faced with dynamic, complex, and potentially life-threatening situations.

Response Generation and Selection

Researchers have investigated response-option generation and selection from a variety of perspectives. This section provides an overview of three bodies of work, each of which informs our understanding of response option generation and/or selection in

different ways. First, research in dynamic domains, such as sport, have investigated the visual search behaviors and domain-specific knowledge representations that are associated with superior response option selection. Second, similar to the studies on hypothesis generation in ill-defined problems (e.g., Gettys & Fisher, 1979), researchers have used text-based problems to investigate the number, and utility, of options generated to solve a problem. Third, researchers in complex and/or dynamic domains (e.g., chess, aviation, sport) have examined, from the perspective of the RPD model, the response option generation and selection strategies employed by skilled performers. The goal of this section is to highlight key methods and findings that contribute to our understanding of response option generation and selection.

Response Generation/Selection Tasks to Examine Perceptual-Cognitive Expertise

Using dynamic tasks, a number of studies have examined the perceptual and cognitive foundations of decision making by presenting participants with representative task stimuli and asking them to enact responses. Typically, these studies have focused on the quality of the selected response, and not on the response options that the individual may have considered prior to selection. For example, Helsen and Pauwels (1993) used a response generation/selection task to investigate the relationship between performance and visual search patterns in soccer. Fifteen skilled soccer players with at least 10 years of competition experience, and 15 novice players with limited competition experience, observed video clips of soccer play from the perspective of one of the on-field attackers. The clips were projected onto a solid wall; participants wore a head-mounted eye tracker and stood with a soccer ball at their feet. At a point in each clip, an attacking player in the video passed the ball toward the camera, and the participant enacted a response by

either: (a) kicking the ball at the goal, (b) dribbling the ball to simulate moving around a defender, or (c) passing the ball toward an on-screen team mate. A panel of soccer experts determined the correct response choice (i.e., shoot, dribble, or pass) for each clip. An analysis of the data revealed that the skilled players selected the correct response significantly more frequently ($M = 91.76\%$) than the novice players ($M = 82.23\%$). Additionally, the skilled players were significantly faster in initiating movement and making foot/ball contact compared with the novice players. Analysis of the eye movement data for passing situations revealed significant skill-based differences in the number and location of fixations: skilled players made fewer fixations than the novice players, and primarily fixated the pass receiver, the free defending player, and the free space. Novice players looked at the same cues, but also attended to other, less-relevant cues, including the attacking players, the goal, and the ball. The authors concluded that the skilled players selected more correct responses because they selectively attended to the most meaningful cues, paid little attention to less informative cues, and processed information faster than the novice players.

McPherson & Kernodle (2003) investigated the link between tennis players' response selections, including shot execution, and their underlying knowledge representations. Based on previous research (e.g., McPherson, 1999), McPherson and Kernodle predicted that skilled performers would employ a refined set of rule-based actions to guide their responses, and would use situational information gathered in-event (e.g., opponent's tendencies and strengths/weaknesses) to adapt their responses to the current game situation. Participants were six professional tennis players with tournament experience and six novice players with no tournament experience. Players within each

skill group were paired randomly, and each pair played one tennis match. Following each point in the second set, players provided immediate retrospective verbal reports of their thinking. After completion of the matches, which were videotaped, three tennis experts rated the appropriateness (i.e., high/low quality) of each player's response selection for all serves and shots. Analysis of the response selection data revealed that the professional players made significantly more high-quality selections for serves ($M = 96.5\%$) and for shots ($M = 95.1\%$) than the novice players ($M = 64.7\%$ and $M = 65.0\%$ respectively).

The verbal report data were coded to identify the type, variety, and sophistication of players' concepts. Skilled players' verbalized significantly more situation assessment concepts (i.e., concepts related to the conditions under which they should respond) than novices, and updated these concepts as the match progressed based on their opponent's tendencies, strengths, and weaknesses. They also demonstrated more tactical behavior in that their situation assessment concepts were directly associated with specific responses (i.e., the direction, speed, and placement of the participant's own responses). On the other hand, novices verbalized situation assessment concepts that focused on aspects of their own (rather than their opponent's) play, and their response concepts often lacked specificity. McPherson and Kernodle (2003) concluded that skilled players' ability to select high-quality responses was supported by a conceptual knowledge base that was more varied, sophisticated, and interrelated than that used by novices.

The studies by Helsen and Pauwels (1993) and McPherson and Kernodle (2003) are representative of a large body of research in sport that has highlighted the link between perceptual-cognitive skills and skilled performance. As in the case of the two studies reviewed here, research in sport that has assessed participants' ability to select

and execute response actions has typically focused on the appropriateness, or quality, of selected responses. Measures of response selection quality have primarily been used to provide evidence of skill-based differences in performance which, in turn, provided a basis for comparing process measures, such as visual search characteristics and in-task cognitions. However, while such research has employed innovative methods to elicit and assess response selections, such as requiring participants to enact responses during actual and simulated task performance, it has rarely addressed the option generation and selection strategies used by skilled performers (but see Johnson & Raab, 2003; Raab & Johnson, 2007).

In other domains, research that has investigated, specifically, aspects of response option generation has done so from two main perspectives. One perspective suggests that in order to select a high-quality response option, the decision maker must first generate most, if not all, of the potential response options (e.g., Engelmann & Gettys, 1985). Studies of response option generation from this “more is better” perspective have employed text-based descriptions of ill-defined problems to examine the number, breadth, and utility of response options generated as potential problem solutions. A different perspective, based on one of the main findings of the RPD model, suggests that skilled decision makers can select a satisfactory, if not a very good, response option as the first one they consider (e.g., Klein, 1989). According to the latter perspective, there is no advantage to be gained by generating a large number of alternate response options. Studies from this “less is better” perspective have investigated the response-option generation and selection strategies used by skilled performers in a range of complex and dynamic tasks. The next sections review research related to each perspective,

highlighting the methods used to elicit response options and describing findings related to the characteristics of response option generation and selection.

Response Option Generation in Ill-Defined Problems

Similar to the studies that examined hypothesis generation in ill-defined problems (e.g., Gettys & Fisher, 1979; Mehle, 1982), researchers have also investigated the generation of response options (i.e., potential solutions) to problem situations. For example, Engelmann and Gettys (1985) associated superior option generation performance with the ability to generate all possible responses, and examined response option generation and its relation to divergent thinking ability. In one experiment, undergraduate and graduate university students were presented with a fictional problem concerning a foreign student who was attending a US university, but had no money for housing. Participants generated as many response options as possible that the foreign student could take to secure housing, and then estimated the number of reasonable options remaining (i.e., those they had not generated). To assess the completeness of option generation performance, the authors pooled the participants' responses and structured a hierarchical tree that contained nine major option categories and a total of 67 reasonable options (i.e., those with positive utility). On average, the graduate students generated significantly more options ($M = 17.75$) than did the undergraduate students ($M = 10.58$). However, participants in both groups underestimated the number of reasonable options they had not generated; estimates generally ranged between 4–5 options. Additionally, both groups' response options represented only a subset of the nine major option categories ($M_{\text{graduates}} = 5.0$, $M_{\text{undergraduates}} = 3.4$).

Although Englemann and Gettys (1985) regarded participants' response option generation performance as less than ideal, they suggested that a more appropriate measure of performance was the extent to which participants generated the high-utility portions of the hierarchical tree. A subsequent analysis using previously established utility values for the foreign student housing problem revealed that both the undergraduate and graduate students' best (i.e., highest utility) two response options were equal in value to the two highest-utility options in the tree. When option sets of increasing size (i.e., 3–5) were considered, the graduate students generated higher-utility response options than did the undergraduate students. Across all participants, divergent thinking ability accounted for 24% of the variance in response option generation performance, and graduate students scored significantly higher on a test of divergent thinking than did the undergraduates. The authors concluded that while all participants were able to generate at least one high-utility option, failure to generate a complete set of response options could be detrimental in more serious problem situations where selecting a sub-optimal option has serious consequences. Note that although the authors compared students who had different levels of tertiary education—ostensibly to examine the effects of experience/knowledge on option generation—this comparison was not a major focus of the study.

In a separate study using similar, ill-defined problems, Gettys et al. (1987) again found that while all participants generated at least one high-utility response option, they generated relatively few options compared with the total number of pooled options. Gettys et al. hypothesized that participants' failure to generate more complete option sets may have been attributable to a lack of motivation. This notion was tested in a follow-up

experiment, in which sixty undergraduate students were randomly allocated to one of three incentive conditions. In the two treatment conditions, participants were paid based on either the quality or quantity of response options they generated; participants in the control condition received no incentive. In this experiment, participants generated response options to solve their university's shortage of car parking spaces, and then estimated the number of options they had not generated. The responses were pooled and then organized into a hierarchical tree with seven major option categories. Analysis of the data indicated that incentive condition did not significantly affect the quantity or quality of options generated; all participants generated approximately 8–9 options, including at least one high-utility option. Across all three conditions, participants generated an average of 3.7 of the seven major option categories. Additionally, as in the Engelmann and Gettys (1985) study, participants overestimated the completeness of their option sets.

Although Engelmann and Gettys (1985) and Gettys et al. (1987) regarded participants' inability to generate near-complete sets of response options as problematic, their findings actually suggest that it is not necessary to generate all possible options in order to generate a good (i.e., high quality) option. Note, however, that these studies did not address the relationship between the quality of response options and their order of generation. Participants may have generated relatively few response options because those options were generated in order of descending quality. If the decision maker's goal is to generate a high-quality option and if higher-quality options are generated before lower-quality options, then there is little reason to engage in extensive response option generation—the first option generated should be a high-quality option. Recall that this notion is one of the main claims of the RPD model with regard to making decisions about

courses of action: experts' superior ability to recognize situations allows them to generate a satisfactory, if not a good, course of action as the first one considered—thereby obviating the need for extensive option generation. This view is also consistent with Simon's (1956, 1990) notion of bounded rationality, which gives rise to satisficing (cf. optimizing) behavior due to limitations in the decision maker's ability to process information (e.g., fallible memory, limited capacity) and the ecological constraints on decision making (e.g., limited time, the cost of acquiring additional information). The next section reviews studies that have tested aspects of this claim in a variety of complex and dynamic domains.

Empirical Studies of Response Option Generation in Complex and Dynamic Domains

Early studies that applied the RPD model examined the frequency with which decision makers generated a single, good response option. For example, Calderwood, Crandall, and Klein (1987) conducted critical decision method (CDM) interviews to examine the response-option generation and selection strategies employed by experienced urban firefighters. Twelve firefighters with substantial command experience (i.e., experts) and 12 firefighters with limited command experience (i.e., novices) were asked to recount a challenging fire incident and identify the points at which they made decisions. For each decision point, the researchers used a series of probe questions to ascertain the number of response options generated, and in cases where multiple options were generated, whether commanders evaluated those options serially or concurrently. Analysis of the interview protocols revealed that both the experienced and novice commanders generated a single response option in approximately 70% of decision points.

When multiple response options were considered, the experienced commanders employed both serial (18%) and concurrent (12%) evaluation, while novice commanders employed only a concurrent evaluation strategy (30%). The authors concluded that in dynamic and complex domains, skilled decision makers typically select the first response option they generate, and that expert decision makers rarely engage in the direct comparison of multiple alternatives. However, because each participant reported on a different critical incident, it is possible that the number of options generated may have been dependent on the characteristics of the situation (e.g., complexity).

In one of few experimental tests of the RPD model, albeit in a relatively static domain, Klein, Wolf, Militello, and Zsombok (1995) used an option-generation paradigm to compare medium- and highly-skilled chess players' ability to generate and select the next best move. Participants viewed a chess board configuration from a real game, generated all possible next moves they could take, and then selected their final move from those generated. Based on the ratings of Grand Master chess players, all chess players considered acceptable moves first, generally chose as their final move one of the moves that were generated earlier in the sequence of options, and generated only a relatively small number (i.e., 3–5) of the possible legal moves. The key point is that for the more experienced players, the first option generated was often a very good option. However, skill groups did not differ much in their option-generation behavior.

Stokes, Kemper, and Kite (1997) investigated the relationship between pilots' domain-specific knowledge and their decision strategies. Twelve expert pilots (i.e., with > 1,500 hours' flight time) and 12 novice pilots (i.e., with < 50 hours' flight time) observed, but did not interact with, a computerized instrument panel that displayed a pre-

programmed flight scenario. The scenario consisted of periods of routine, incident-free flight interspersed with problem situations. In a subset of the problem situations, the pilots were prompted, on-screen, to (a) list all relevant cues, (b) enter all plausible response options, and then (c) select the best response option. Expert raters (i.e., high flight-time pilots) evaluated the quality of participants' selected response options, and determined cue relevancy. An analysis of the data revealed that the expert pilots generated 30% more response options than the novices; they also selected response options that were significantly higher in quality than those selected by the novices. When pilots generated multiple response options, the expert pilots selected their first-listed option more frequently than did the novices (71% versus 53% of cases, respectively). The expert pilots also listed significantly more ($M = 15.4$) relevant cues than the novices ($M = 8.9$), and although both groups listed similar numbers of irrelevant cues, the expert pilots listed proportionally fewer (39.1%) irrelevant cues than the novices (59.0%). Further analysis revealed that the best predictor of decision-making performance was the number of relevant cues listed. The findings of this study provide mixed support for the RPD model: while the model would predict that as expertise increases the number of response options generated should decrease, the expert pilots actually generated more response options than the novices. However, in support of the RPD model, the expert pilots selected their first-generated response option in a majority of cases. The fact that the expert pilots reported more relevant cues and proportionally fewer irrelevant cues than the novices suggests that a more detailed and accurate understanding of the situation facilitates the generation of a range of response options.

Johnson and Raab (2003) used a similar option-generation paradigm to that employed by Klein et al. (1995) and Stokes et al. (1997) to assess decision-making strategies in a dynamic task. Their research was designed to test specific predictions of the Take-The-First heuristic, which is built on the premise that in some dynamic, uncertain, time-constrained situations (e.g., invasion sports)—where there is little opportunity to deliberate between options—generating fewer, rather than more, options results in better decisions. This heuristic was developed specifically based upon empirical data from the study of skilled performance in handball; the implication was that it may extend to other similar environments. Take-The-First heuristic—like other fast and frugal heuristics—emphasizes the adaptive nature of decision making, and assumes that use of the heuristic depends on the fit between the structure of the environment and the nature of the decision maker (see Gigerenzer & Brighton, 2009; Gigerenzer, Todd, & the ABC Research Group, 1999; Goldstein & Gigerenzer, 2002). Consistent with the RPD model, Johnson and Raab suggested that individuals should generate only few options and then select one of the initial options as the best option because of a hypothesized negative relationship between the number of options generated and the quality of a decision. According to advocates of the Take-The-First heuristic, not adopting an RPD-type strategy (i.e., selecting the first option generated) would increase the likelihood of making a poorer decision.

In their study, Johnson and Raab (2003) asked moderately-skilled handball players to watch, from the perspective of the attacking team, 10-second video clips of high-level handball games that ended with an attacking player in possession of the ball in front of the goal. The final frame was frozen on-screen for 45 seconds while participants

imagined themselves as the player with the ball and: (a) named, as quickly as possible, the first decision that intuitively came to mind; (b) named as many additional options they could conceive; and (c) selected the best option from those they had generated. Expert coaches rated the participants' options for each clip using a 5-point scale. On average, players generated between 2–3 response options. Analysis of the data supported the hypothesis that decision quality was negatively correlated with the number of options generated. When players generated fewer options they were more likely to choose a better option as their final choice. Moreover, the first option generated was rated of higher quality than the second option, and the second option was rated of higher quality than the third, etc. Johnson and Raab's original hypotheses were supported by data from a subsequent study of the option-generation strategies of skilled and less-skilled handball players (Raab & Johnson, 2007). The effect of skill was similar to that observed by Klein and Peio (1989): skilled players generated the highest-quality move as their first option, and selected that option as their final option more frequently than did the less-skilled players.

Moxley, Ericsson, Charness, and Krampe (2012) examined whether chess players' decisions are based solely on fast, intuitive processes (e.g., TTF) or whether move quality is improved after a period of slower deliberation. Thirty-seven expert chess players ($M_{\text{ELO rating}} = 2194$, $SD = 130$) and 34 tournament players ($M_{\text{ELO rating}} = 1836$, $SD = 92$) thought aloud while selecting the next move for 15 chess problems. The analysis focused on the six problems (2 easy, 2, medium, 2 hard) for which there was a clearly defined best move. The authors hypothesized that if players' decisions were based on intuitive processes, there would be little difference in move quality between the first,

intuitive move mentioned and the final move selected after a period of slow deliberation. Alternatively, if mechanisms that support slower, deliberative processing—such as LTWM—are an important component of skilled decision making in chess, the final move selected should be of higher quality than the first move mentioned. The analysis revealed that, as expected, the expert players' first-mentioned and finally-selected moves were of higher quality than those of the tournament players. However, both groups benefitted from additional deliberation time: their final (i.e., selected) moves were of higher quality than their first-mentioned moves. The improvement in move quality after deliberation was evident at each level of problem difficulty, but was greatest for the hard problems (Cohen's $d = 0.94$). Additionally, problem difficulty affected players' choice of final move: both groups were significantly more likely to select their first-mentioned move as their final move for the easy problems, than for the hard problems. The authors concluded that although experts did make better intuitive decisions than less-skilled players, engaging in slower, deliberative processing resulted in improved move quality, regardless of skill level and problem difficulty.

To summarize, research that has investigated response option generation and selection has found that, regardless of skill level, performers typically (a) generate a small number of options, (b) generate at least one good option, (c) generate higher-quality options earlier in a sequence than lower-quality options, and (d) select their first-generated option. When skill-based differences have been observed in time-constrained situations, skilled performers have generally selected higher-quality options and selected their first-generated option more frequently than less-skilled performers. In some studies, accurate response selection was found to be related to individuals' ability to prioritize

task-relevant cues while disregarding task-irrelevant cues, assess their opponent's tendencies and actions, and associate responses with specific conditions—all of which are factors associated with situation assessment. While this supports a close relationship between the situation assessment and response phases of decision making, relatively few studies have investigated, empirically, the relationship between the generation/selection of options in the situation assessment and response phases of decision making. The next section reviews research that has examined this relationship.

Option Generation/Selection in the Situation Assessment and Response Phases of Decision Making

A comprehensive search of the literature revealed only two studies that have investigated the relationship between the situation assessment and response phases of decision making by focusing on the generation and selection of options. Note that both studies did not elicit options directly—rather, options were extracted from either concurrent or retrospective verbal reports.

Using the campus car parking problem (see Engelmann & Gettys, 1985), Adelman, Gualtieri, and Stanford (1995) investigated the relationship between the situation assessment and response phases of decision making by varying the information contained in the problem description. Twenty graduate students were randomly assigned to one of four causal conditions. In the control condition, the problem description included only basic information about the car parking situation on a university campus. The other three causal conditions contained additional information about either: (a) rapid growth of the university's population, (b) inefficient use of existing car parking spaces, or (c) both rapid growth and inefficient usage. In each condition, participants thought aloud

as they considered the problem and then presented their recommended course of action. An analysis of the think-aloud data revealed that the causal information affected the types of options that participants generated and selected in both the situation assessment and response phases of decision making. For example, participants who received information about the inefficient use of existing car parking spaces generated proportionally more response options designed to improve efficiency than those who received information about rapid growth or those in the control condition. However, the authors noted that the results were not as strong as expected, and attributed this, in part, to participants' relative lack of experience with the problem situation. They subsequently suggested that expertise in a given domain may affect the strength of the relationship between option generation/selection in situation assessment and response.

Ward et al. (2011) examined the option generation strategies employed by police officers during complex and dynamic situations. Rookie and experienced police officers were equipped with a replica handgun and interacted with video simulations of high- and low-frequency law enforcement incidents, some of which required a lethal (i.e., shoot response). Successful performance on the lethal trials required officers to perceptually anticipate and act upon the eventual situational outcome (e.g., a suspect who draws a gun and shoots at the officer). Officers provided retrospective verbal reports about their reasoning after several key trials. Based on Long Term Working Memory theory (LTWM; Ericsson & Kintsch, 1995), the authors expected that successful performance would be supported by a detailed situation model that would include a range of situational outcomes. Therefore, they predicted that successful performers would generate more task-relevant situation assessment options than less-successful performers.

They also predicted that having a more detailed understanding of what could happen during situation assessment would allow officers to adopt a TTF-like strategy (i.e., generate relatively few options) during the response phase of decision making. Consequently, they predicted a negative relationship between the number of options generated in the two decision phases.

Following the experiment, the verbal report data for three lethal trials were segmented and coded as either situation assessment or response options. Contrary to expectation, an analysis of the data revealed that the number of options generated during situation assessment was positively related to the number generated during the response phase of decision making. Additionally, the number of options generated in the situation assessment and response phases were both positively related to successful performance. Furthermore, compared with the rookies, experienced officers generated a significantly greater number of task-relevant situation assessment and response options. The authors concluded that an option generation strategy based on LTWM theory (i.e., generating more, rather than fewer, relevant options) could support option generation in both the situation assessment and response phases of decision making.

Although these studies provide some preliminary evidence about the relationship between the situation assessment and responses phases of decision making, use of an option-generation paradigm—rather than inferring options from verbal report data—might permit a greater understanding of the mechanisms supporting superior performance. Moreover, an option-generation paradigm would allow some of the current and most dominant theoretical descriptions of expertise (e.g., LTWM, RPD, Take-The-First) to be contrasted experimentally.

Summary

In summary, this section considered ways in which response generation and selection tasks have been used to investigate skilled performance. Studies that used response selection tasks in dynamic domains primarily established that superior response selection is driven by perceptual-cognitive skill. Studies that employed response generation tasks to investigate performance in ill-defined problems demonstrated the utility of response option generation tasks and established that people tend to generate relatively few of the plausible response options—although their generated options typically include at least one high quality response. The empirical studies of response option generation in complex and dynamic domains highlighted theoretical perspectives on option generation. Both the RPD and TTF approaches suggest that individuals generate few response options, and that better options are generated first. According to these perspectives, skilled decision makers should tend toward generating a single, good option. Finally, it was demonstrated that option generation tasks can be used to investigate the situation assessment and response phases of decision making simultaneously. One study, from the law enforcement domain, found evidence that that contradicts the RPD/TTF account of option generation; skilled performers generated more, rather than fewer, task-relevant options in both phases of decision making. The next chapter summarizes the main hypotheses that can be drawn from LTWM theory, RPD, and TTF, and provides a brief outline of the studies described in Chapters 5 and 6.

CHAPTER 4

GENERAL HYPOTHESES AND OUTLINE OF STUDIES 1 AND 2

General Hypotheses

Based on the literature review, two perspectives on option generation can be identified: LTWM theory, which focuses on the situation assessment phase of decision making, and RPD and TTF, both of which focus on the response selection phase of decision making.

In general, LTWM theory posits that superior performers generate a detailed, up-to-date, and accurate representation of the situation, and that this facilitates their ability to anticipate the situational outcome. Based on the Ericsson and Kintsch's (1995) original description of LTWM, Ward et al. (2012) developed quantitative hypotheses related to option-generation behavior. Ward et al. asserted that when multiple, task-relevant options are present in the environment, more of those options will be encoded in the representations of superior performers, compared to lower-level performers' representations. Consequently, Ward et al. hypothesized that superior performers not only could, but would, generate more task-relevant options than lower-level performers. Furthermore, as superior performers have a refined ability to anticipate situational outcomes, Ward et al. expected that the number of task-relevant options generated would be positively correlated with prediction quality. Ward et al. suggested that the relationship between task-relevant assessment options—in terms of the relative threat associated with each option—would form an integral part of the information encoded in the situation model. Therefore, superior performers should be better than lower-level

performers at prioritizing task-relevant assessment options in terms of the threat posed by each option. The data from the Ward et al. study supported these hypotheses. Henceforth, reference to LTWM hypotheses will relate to the hypotheses outlined by Ward et al.

As a corollary, it is possible that generating more task-relevant options during assessment—painting a detailed picture of the situation—could facilitate a TTF-like mechanism during response selection. That is, the ability to anticipate the situational outcome could enable experienced decision makers to generate, as their first response option, a satisfactory, if not a very good, response, without the need for generating more. Such a finding would permit some degree of complementarity between LTWM and TTF—both could be at play, albeit in different phases of decision making. As noted, however, preliminary evidence from Ward et al. (2011) did not find support for this hypothesis. I explore this further in this dissertation using an option generation paradigm.

With respect to the response phase of decision making, both RPD and TTF suggest that individuals generate fewer, rather than more (cf. LTWM theory) response options, and then select one of the first options generated. To explain this finding, TTF posits that the number of response options generated is negatively correlated with the quality of the final decision—because better options are generated first. However, while LTWM theory makes the distinction between task-relevant and task-irrelevant options (i.e., during situation assessment), neither RPD or TTF make such a distinction. For instance, the negative relationship hypothesized by TTF is between the total number of response options generated, and the quality of the final decision.

Outline of Studies 1 And 2

Collectively, the purpose of Studies 1 and 2 is to extend the work of Ward et al.

(2011) by using a prediction and option-generation paradigm to investigate situation assessment and response selection in dynamic and complex law enforcement situations. Study 1, which is methodological in nature, employs converging approaches to identify video-based law enforcement scenarios that could discriminate between experienced and less-experienced police officers, in terms of their ability to anticipate the culminating event. Study 1 also serves to identify a criterion occlusion point in each discriminating scenario: the point that maximized the experienced-based difference in anticipation performance. Study 2 then employs those discriminating scenarios to test specific hypotheses derived from LTWM theory and TTF (and to a lesser extent, RPD) that relate to outcome measures of assessment and response selection performance, and the option generation processes that underlie superior performance.

CHAPTER 5

STUDY 1

Perceptual anticipation—the ability to utilize environmental information cues to anticipate how a situation may unfold in the immediate future—is a hallmark of expertise (Suss & Ward, in press). One method that has been employed to investigate perceptual anticipation is the temporal occlusion paradigm. In video-based temporal occlusion studies, participants typically view video stimuli depicting dynamic action sequences (e.g., an opposing tennis player’s serve, a developing play in soccer). The stimuli are occluded unexpectedly (i.e., the screen turns black) at a specific point(s) during the developing sequence, and participants are asked to predict the outcome of the situation (e.g., the future location of an opponent’s tennis serve; the intention of a soccer player—shoot at goal, pass to another player, or continue running with the ball).

Temporal occlusion studies typically seek to identify the time point, or time window, during which experts demonstrate an advantage in their ability to use perceptual information to correctly anticipate the outcome. In sport-related studies that have employed the temporal occlusion paradigm, occlusion point(s) have often been determined by objective ecological events, such as the point of ball-racket contact in racket sports (e.g., Abernethy & Russell, 1987), foot-ball contact in soccer (e.g., Savelsbergh, Onrust, Rouwenhorst, & Van Der Kamp, 2006; Ward et al., 2012), ball release in cricket (e.g., Mann, Abernethy, Farrow, Davis, & Spratford, 2010), or weapon impact in fencing (e.g., Hagemann, Schorer, Canal-Bruland, Lotz, & Strauss, 2010). For example, Ward et al. (2012) occluded clips of soccer plays approximately 120 ms prior to foot-ball contact. These “events” are typically deterministic and governed by the laws of

physics (e.g., direction and velocity of a “pass” is determined by kinematic and kinetic variables that precede it). In addition, more probabilistic information is available to dis/confirm predictions made from this deterministic information, such as the preceding and/or co-occurring movements and locations of their teammates relative to other opposing players, team mates, and the goal.

In other real-world, dynamic situations—such as law enforcement confrontations—physically deterministic events are often less informative with regards to the ultimate outcome of any given action. For instance, deterministic perceptual cues may be available to tell the observer that a perpetrator is reaching in to his/her pocket, but not what the perpetrator will take out (e.g., a weapon versus driver’s license). Likewise, in sport, the number of possible outcomes for any given action is often highly constrained by the rules of the game (e.g., shoot, pass, retain possession of the ball) whereas in complex domains that have fewer constraints the possible outcome of any given action is relatively unlimited. Consequently, observers have to rely more on probabilistic information cues to predict the outcome. Accordingly, to measure participants’ perceptual anticipation skill, the point of occlusion has to be based on these probabilistic events.

The overarching goals of this study were methodological: (i) identify dynamic law enforcement scenarios that could discriminate between experienced and less-experienced police officers’ ability to perceptually anticipate the outcome, and (ii) determine the optimal occlusion point that maximizes the skill-based difference in anticipation. Rather than choose occlusion points for each scenario at random, Studies 1A, B, and C were conducted to generate empirical data that converged on specific ecological cues and occlusion points. These data allowed for the generation of some

preliminary hypotheses that the identified occlusion points could be used to differentiate between skill groups in terms of their ability to predict the outcome. Accordingly, I first present the method employed in each study and then present the combined analysis of all three studies together.

STUDY 1A

In this preliminary study, a single occlusion point—selected by the experimenter—was assessed for each scenario. In addition to the main goal of assessing the occlusion points, a secondary goal was to test the experimental tasks and procedures that would be used in Study 2 to investigate option-generation during the assessment- and response-selection phases of decision making. Therefore, although the primary focus of Study 1A was on participants' ability to anticipate the outcome of each scenario—which requires participants to generate only a single option (i.e., the anticipated outcome)—the experimental tasks were designed to elicit this information using an option-generation paradigm (partly as a means to refine it for use in Study 2). The option-generation paradigm differs from an anticipation paradigm (as used in Studies 1B and 1C) in four important ways: (i) instead of generating a single anticipation (i.e., what they think *will* occur next), participants can generate multiple assessment options (i.e., what they think *could* occur next); (ii) participants are asked to rate each option in terms of its likelihood of occurring, (iii) the anticipated outcome is assumed to be the assessment option rated highest in likelihood; and (iv) in addition to generating assessment options (e.g., courses of actions that could be taken by the perpetrator or others in the scenario), participants are also asked to generate response options (e.g., personal courses of action that could be enacted by a police officer in response to a perpetrator's actions).

Method – Study 1A

Participants

Experienced participants were 11 male, SWAT-qualified law enforcement officers with a median of 9 years of experience (range: 6–18 years) and an average of 2082 hours ($SD = 724$) of accrued training time. Fourteen *less-experienced* participants were male law enforcement recruits who were near completion of a 21-week training program. Twelve of the less-experienced participants had logged an average of 174 ($SD = 48$) hours observing police officers on patrol (i.e., during “ride-alongs”); the remaining two had not logged any observation time. Three of the less-experienced participants—including the two who had not logged any observation time—had served in the military; two served in reserve units for approximately three years, and one served full-time for 4.5 years.

Materials and Stimuli

Twenty interactive, video-based law enforcement scenarios (IES Interactive Training: Ann Arbor, MI) served as the basis for the experimental stimuli (see Appendix A). The scenarios, ranging from 19.34–93.88 seconds in length, depicted high and low frequency events encountered by officers in the line of duty (e.g., disturbance call, potential suicide, domestic assault, traffic stop, suicide bomber; see Appendix B for exemplar scenarios). Fourteen *escalating* scenarios culminated in the suspect simulating an attack (e.g., swinging a baseball bat, shooting firearm, striking with a rock, detonating an improvised explosive device). Six *de-escalating* scenarios culminated in the suspect(s) surrendering or calming down.

To create stimuli that allowed participants to generate assessment and response options, a critical *turning point* for each clip was first identified. A turning point was defined as a change in the dynamics of the unfolding events that is marked by a positive shift in ecological validity (i.e., when task-relevant contextual information becomes available that provides a basis for predicting the outcome of the scenario) (Suss & Ward, in press). For example, in the Blow Up scenario (see Appendix C), the critical turning point occurred when the boy bent down to pick up the bomb's actuator from the ground (i.e., after he had dropped it, but prior to detonating the device). In each scenario, the turning point typically occurred within three seconds of, and presaged, the culminating event.

Next, a specific occlusion point was identified—the earliest point where sufficient information was available to allow experienced participants to accurately predict the culminating event. For example, in the Blow Up scenario, the point at which the boy raised his hand to strike the bomb's actuator (i.e., after picking up the actuator from the ground) was selected as the occlusion point. Identification of the occlusion point in each scenario was based on (a) formal and informal reviews of performance and retrospective verbal report data from previous research that employed the same scenarios (Tashman et al., 2006; Ward, Harris, Ericsson, Eccles, & Tashman, 2007; Ward et al., 2011), (b) pilot testing of the experimental tasks which was conducted with 10 law enforcement officers, (c) conversations with several law enforcement trainers, and (d) personal intuition, based on previous military and security experience.

After the occlusion point was identified, each scenario was digitally edited to end at that point (i.e., was replaced by black screen), thus creating the video stimuli. Each

video stimulus began with a brief (i.e., approximately 10 s), audio-only “radio message” describing the situation to which the officer was being dispatched (see Appendix E). Each video stimulus was inserted into a Microsoft PowerPoint 2010 presentation, which also included prompts for the subsequent assessment and response option generation tasks.

Option Generation Tasks

Assessment option generation task. In this task, participants were required to generate assessment options—events that could occur in the environment (e.g., courses of actions that could be taken by the perpetrator or others in the scenario) at the point of occlusion. To elicit assessment options, participants responded verbally to the following textual prompt, which was displayed on the screen: “What could happen next on the screen in the next few seconds?” The prompt was displayed together with a 10-second countdown timer and a single text box. Participants were required to verbalize their first option before the timer reached zero; the experimenter transcribed the option into the text box. This process was repeated using a new slide/timer for each additional option, until the participant indicated that they had verbalized all of the assessment options that were heeded at the point of occlusion.

To ensure that participants reported only those options heeded at the point of occlusion, and to prevent them from deliberating about options while completing the option generation tasks, preparatory instructions were provided during a practice trial. Specifically, participants were instructed to (a) list, rapidly, only those assessment options that they were considering at the point of occlusion; (b) verbalize all of the assessment options under consideration, even if participants believed a specific option(s) to have a low likelihood; and (c) refrain from engaging in reflective option generation

(e.g., reporting options that were generated only after reflecting on the scenario, generating options solely for the sake of increasing the number of options generated, etc.)

After the participant finished verbalizing their assessment options, the rating tasks were administered. To facilitate rating, the complete set of transcribed options was displayed on a single screen; a small text box next to each option was used to record the ratings. Participants first rated the likelihood that each of their generated options would actually happen after the occlusion point, using a scale that ranged from 0 (*not at all likely*) to 100 (*very likely*). To prevent a participant's likelihood ratings being subadditive or superadditive, the likelihood ratings were required to sum to 100. To assist participants in this task, a "remainder" box at the bottom of the screen displayed the number of to-be-allocated likelihood points; this number was adjusted automatically each time the experimenter transcribed a likelihood rating next to its corresponding option.

Next, using a new screen with blank rating boxes, participants rated how threatening each option was to their own personal safety, using a scale from 0 (*not at all threatening*) to 100 (*very threatening*). As participants may have generated more than one option that they regarded as very low/high threat, they were instructed to assign threat ratings for each option independently (i.e., without regard to the sum of the individual threat ratings).

The set of assessment options displayed on the likelihood- and threat-rating screens were presented in a shuffled order, and not in their order of generation. This step was implemented to reduce the possibility of participants simply rating options according to their serial position on-screen (e.g., assigning a high likelihood to the first option

generated, and a lower likelihood to the next option, etc.). Note, however, that the serial order of option generation was recorded for later analysis.

Response option generation task. Participants completed a similarly structured response option generation task by answering the question, “How could you respond in the next few seconds?” Participants were reminded to follow the preparatory instructions (e.g., report only those options heeded at the occlusion point). After verbalizing their response options, participants first rated their likelihood of pursuing each option, employing the same likelihood rating scale and procedure described for the assessment option generation task. Next, they rated the quality of each option independently, using a scale that ranged from 0 (*not at all good for my own/innocent bystanders’ immediate safety and survival*) to 100 (*very good for my own/innocent bystanders’ immediate safety and survival*).

Procedure

After providing their written informed consent, participants completed a biographical questionnaire, including questions about their law enforcement training, qualifications, and employment, as well as the number of times they had participated in law enforcement training using video-based simulations. Participants wore their standard duty uniform and equipment belt, and were equipped with a replica handgun, inert Taser, inert chemical spray, real baton, and a radio. Participants also wore a lapel microphone (Sony WCS-999) that was connected wirelessly to a video camera (Canon VIXIA HF M300); the camera recorded the participant’s physical actions and verbal communication, as well as the projected video stimuli.

The video stimuli were displayed on a large projector screen (experienced participants: 212 cm x 157 cm; less-experienced participants: 245 cm x 192 cm). Participants began each trial standing directly in front of the screen (experienced participants: 330 cm from screen; less-experienced participants: 520 cm from screen), but were free to move from this position once the video started. The experimenter instructed participants to interact with each video stimulus as if they had been dispatched to the situation while on patrol (e.g., observe, issue verbal commands, move, communicate via radio, deploy the appropriate tactical force option). Participants then interacted with one practice trial; after the occlusion point, the experimenter introduced the assessment and response option-generation tasks. Specifically, the experimenter instructed participants to: (a) list, rapidly, only those assessment/response options that they were considering at the point of occlusion; (b) verbalize all of the assessment/response options under consideration, even if participants believed a specific option(s) to have a low likelihood; and (c) refrain from engaging in reflective option generation (e.g., reporting options that were generated only after reflecting on the scenario, generating options solely for the sake of increasing the number of options generated, etc.). Participants then completed the assessment and response option-generation tasks for the practice trial.

After the practice trial, participants completed the 19 test trials in a randomized order. Participants always completed the assessment option-generation task followed by the response option-generation task. After completing the two option-generation tasks in a given trial, participants indicated whether they had seen that scenario with/out that particular ending previously in their law enforcement training. Participants did not receive feedback on their performance (i.e., they were never told nor shown how the

scenario culminated after the occlusion point). After each trial, the data (i.e., options and ratings) were automatically written to a text file for later extraction and analysis.

Participants were tested individually; each testing session lasted approximately 2 hours. At the conclusion of each testing session, the experimenter debriefed the participant and thanked them for their participation.

Data Preparation

For each scenario, the *anticipated outcome* was the assessment option that participants rated highest in likelihood. Note that as participants were not instructed specifically to provide unique likelihood ratings, there were some cases in which they assigned the highest likelihood rating to more than one option (e.g., 50/50, 20/40/40, 33.33/33.33/33.33). In such cases, the anticipated outcome was deemed the assessment option with the highest likelihood rating that was generated earliest in the generation sequence.

STUDY 1B

In Study 1A, the experimenter selected the occlusion point for each scenario based on extensive pilot testing. In Study 1B, the occlusion points were identified via a cognitive task analysis conducted with experienced police officers. I will first describe the cognitive task analysis, and then the subsequent study.

Method – Cognitive Task Analysis

Participants

The participants were four experienced police trainers from a large police department. The trainers had a median of 21.25 years' (range: 21.00–25.25) law enforcement experience.

Materials and Stimuli

Seventeen of the 20 interactive, video-based law enforcement scenarios used in Study 1A were used for the cognitive task analysis; three additional escalating scenarios were also included (see Appendix A). The video scenarios were presented in their entirety and were not temporally occluded at any point.

Procedure

The video scenarios, with a frame rate of 29 frames per second, were played using KMPlayer software (<http://www.kmpmedia.net/>) and projected onto a screen (267 cm x 149 cm) using an Optoma TW1692 projector. The trainers observed the scenarios passively (i.e., without interacting) from a seated position, 4.13 m from the screen. The experimenter instructed the trainers that their goal was to identify the earliest point in each video scenario at which they believed experienced law enforcement officers should be able to correctly anticipate the outcome. The trainers first observed a video scenario in its entirety, played at normal speed. Subsequently, the trainers directed the experimenter to replay the video—using the slow motion and pause functions where necessary—until the trainers determined the optimal occlusion point. The trainers were allowed unlimited time to scrutinize the videos, and could re-watch the videos until they were satisfied with their selected occlusion point. The experimenter recorded the exact time point/frame number that corresponded with the selected occlusion point. Each trainer was consulted individually; depending on the trainer, their review of the videos lasted between 1 and 4 hours. Due to time constraints, each trainer viewed a subset of the 20 scenarios (see Table 1). Sixteen scenarios were viewed by at least two trainers.

Table 1
Occlusion Points Selected by Experienced Police Trainers in Study 1B (Cognitive Task Analysis)

Scenario	Time (s) of selected occlusion point				
	Trainer 1	Trainer 2	Trainer 3	Trainer 4	Experimenter
Bar ^a	56.79	56.72	56.90	57.55	57.59
Blow up ^a	33.07	33.24	33.79	30.90	33.17
Blue dodge	30.62	–	30.66	30.72	30.62
Bus lady	–	–	–	–	–
Convenience store	39.62	–	39.62	39.66	39.76
Domestic baby ^a	61.10	61.10	61.17	61.17	61.10
Domestic punch ^a	28.79	30.48	28.66	28.76	28.72
Drunk	–	42.79	–	–	–
Greenwood park	–	8.41	10.97	11.07	10.93
Gun grab ^a	41.79	39.72	41.52	40.76	44.79
Hey Mr Wilson	–	63.07	62.90	66.07	66.03
Hospital incident	40.90	–	–	–	–
Hotel	58.55	–	–	–	–
Larceny	–	–	–	–	–
Men in black ^a	30.79	28.93	30.55	30.69	30.55
Miami	–	–	–	–	–
Rock n roll ^a	29.24	29.62	30.45	32.83	29.31
School hostage ^a	38.31	38.03	38.10	38.24	38.24
School shooting ^a	49.07	49.03	49.14	49.14	49.07
Subway ^a	32.48	32.76	32.72	32.24	32.45
Suicide by cop	73.34	–	76.62	–	–
Suicide waitress	–	59.76	59.93	59.72	60.17
Video store	35.52	–	–	–	–

Note. Videos were screened at 29 frames per second; each video frame was approximately 0.03 s in duration. Dashes indicate that a scenario was not evaluated by a trainer.

^aScenarios included in the ICC analysis.

Reliability Analysis – Cognitive Task Analysis Data

An analysis of inter-rater reliability was conducted on the 10 scenarios that were rated by all four trainers (see Table 1). The average length of the 10 scenarios included in the analysis was 42.62 s ($SD = 10.84$). Across the 10 scenarios, the average difference

between the earliest and latest occlusion point identified by the trainers was 1.40 s ($SD = 1.23$). This provides some initial evidence that the four trainers identified a similar point in each scenario as the ideal occlusion point. A more formal assessment of inter-rater agreement across the 10 scenarios was conducted by calculating an intraclass correlation coefficient (ICC; Shrout & Fleiss, 1979). In contrast to inter-rater reliability measures that are calculated based upon the number of agreements/disagreements between the raters (e.g., kappa; Cohen, 1960), an ICC provides a measure of reliability for ratings made on a continuous scale (e.g., time). Using occlusion-point time as the continuous dependent variable, the ICC(2, 4) coefficient was 0.999, with a lower 95% confidence interval bound of 0.997. An ICC of 1.0 represents perfect agreement, and an ICC above 0.8 represents near-perfect agreement.

To determine the extent to which the experimenter's judgment could serve as an additional rating (e.g., for scenarios which were not scrutinized by all four trainers), the experimenter also completed, independently, the same task as the trainers. When the experimenter's data were included in the reliability analysis for the 10 scenarios, the ICC(2, 5) coefficient was also 0.999, with a lower 95% confidence interval bound of 0.997. This result represents near-perfect agreement amongst the trainers' and experimenter's selected occlusion points.

Then, to identify the criterion occlusion point to be used for each scenario in the subsequent study, I selected the median of the trainers' occlusion points. Three of the de-escalating scenarios from Study 1A were not scrutinized by any trainers due to time limitations; as a substitute, the experimenter scrutinized each of these scenarios and

determined the criterion occlusion point. The criterion occlusion point times are listed in Appendix D.

Method – Performance Data

Participants

Forty-two law enforcement officers were recruited to participate in this study. All officers were from the same police department in an urban metropolitan area of western USA. In this department, police academy graduates initially serve as correctional officers in jails, where they often work for several years before being assigned to a regular, street patrol role. Amongst the officers recruited for this study, there was a broad range of experience levels: officers had between 0.12 and 26.02 years' total law enforcement experience, and between 0.00 and 24.43 years' patrol experience. The frequency distributions for participants' *total years of law enforcement experience* and *years of patrol experience* were negatively skewed. To create two main groups that differed substantially in their level of experience, the sample was divided into three groups: experienced, intermediate-experience, and less-experienced using the following criteria: (a) a mean difference of at least 5 years' total law enforcement experience between the experienced and less-experienced groups, (b) maximize the size of the experienced and less-experienced groups, and (c) minimize the size of the intermediate-experience group. The intermediate group ($n = 11$) was not included in the subsequent analysis, and will not be described further.

Experienced participants. Experienced police officers ($n = 16$; 0 females, 16 males) were those who had at least 12 years' total law enforcement experience ($M = 17.96$, $SD = 5.09$, range: 12.34–26.02) and some street patrol experience ($M = 9.37$ years,

$SD = 5.55$, range: 3.25–24.43). At the time of testing, the experienced officers were serving in a variety of roles, including patrol sergeant, detective, and recruit training officer. The mean age of the experienced officers was 42.19 years ($SD = 4.81$, range: 35–55).

Less-experienced participants. Less-experienced officers ($n = 15$; 2 females, 13 males) were those who had less than 5 years' total law enforcement experience ($M = 2.54$, $SD = 1.83$, range: 0.12–4.75) and were still serving as correctional officers in jails (i.e., had no street patrol experience). The mean age of the less-experienced officers was 32.13 years ($SD = 7.11$, range: 24–45).

The remaining 11 officers (3 females, 8 males), who had between 5 and 12 years' total law enforcement experience ($M = 7.09$, $SD = 2.17$, range: 5.17–11.07), completed the experiment but were excluded from the data analysis.

Materials and Stimuli

Twenty-three interactive, video-based law enforcement scenarios (IES Interactive Training: Ann Arbor, MI) served as the basis for the experimental stimuli (see Appendix A). Using the criterion occlusion points determined by the cognitive task analysis, the scenarios were edited to create temporally-occluded stimuli. To prompt participants to verbalize their anticipation, a short, audible tone was digitally added to the video immediately after the occlusion point and a textual prompt (i.e., What happens next on the screen?) was displayed on the screen.

Procedure

After providing their written informed consent, participants completed a biographical questionnaire, including questions about their law enforcement training,

qualifications, and employment, as well as the number of times they had participated in law enforcement training using video-based simulations. Participants wore their standard duty uniform and equipment belt, and were equipped with a replica handgun, inert Taser, inert chemical spray, real baton, and a radio. Participants also wore a lapel microphone (Sony WCS-999) that was connected wirelessly to a video camera (Canon VIXIA HF M300); the camera recorded the participant's physical actions and verbal communication, as well as the projected video stimuli.

The video stimuli were displayed on a projector screen; the testing sessions were conducted at several locations and employed differently-sized screens (minimum screen size: 141cm x 100 cm; maximum screen size: 222 cm x 156 cm). Participants began each trial standing directly in front of the screen (minimum distance: 330 cm; maximum distance: 480 cm), but were free to move from this position once the video stimulus started. The experimenter instructed participants to interact with each video stimulus as if they had been dispatched to the situation while on patrol (e.g., observe, issue verbal commands, move, communicate via radio, deploy the appropriate tactical force option). At the point of occlusion, participants were asked to verbalize what they anticipated would happen next in the video, if it were to continue. Participants then interacted with one practice trial, and practiced verbalizing their anticipation as quickly as possible after the occlusion point. After the practice trial, participants completed the 22 test trials in a randomized order. After completing the anticipation task in a given trial, participants indicated whether they had seen that scenario previously in their law enforcement training. Participants did not receive feedback on their performance (i.e., they were never told, nor shown, how the scenario culminated after the occlusion point). Participants were

tested individually; each testing session lasted approximately 60 minutes. At the conclusion of each testing session, the experimenter debriefed the participant and thanked them for their participation.

Data Preparation

To prepare the data for scoring, participants' verbalized anticipations were transcribed from the video recordings into an Excel spreadsheet.

STUDY 1C

A preliminary analysis of the data from Study 1B revealed that across all scenarios, the experienced group performed significantly better than the less-experienced group. However, across all scenarios the mean percentage of experienced participants who anticipated the outcome correctly was below 50%. Recall that the objective of the trainers' cognitive task analysis was to identify the occlusion point at which all—or at least a majority—of experienced officers should be able to anticipate the outcome correctly. There are at least three explanations for the relatively poor level of anticipation performance demonstrated in Study 1B: (i) some scenarios may not contain cues that are sufficiently salient to facilitate anticipation, (ii) the occlusion points determined by the trainers occurred prior to salient cues becoming available, and (iii) participants' interaction with the video scenarios affected their ability to anticipate the outcome of the scenario.

Therefore, to gain a more detailed understanding of anticipation performance in each scenario, an online, computer-based study was conducted with three occlusion points per scenario. There were two main objectives: to identify which scenarios were

discriminating of experience level, and to determine the optimal occlusion point for those discriminating scenarios. To remove any possible effects of physical interaction on anticipation performance, participants in Study 1C observed the video scenarios passively (i.e., from a seated position).

Method

Participants

Experienced participants. A link to the online study was distributed to experienced law enforcement officers via police training agencies, state tactical officer associations, and law enforcement web sites. Fifty-two experienced law enforcement officers completed the experiment. Complete demographic and law enforcement experience data were obtained for 42 (2 females, 40 males) of the 52 experienced officers. The 42 officers had an average of 14.18 years' ($SD = 8.54$, range: 1–34) law enforcement experience; their average age was 39.57 years ($SD = 10.04$, range: 21–64).

Less-experienced participants. Sixty-five police recruits (9 females, 56 males) were recruited via an informational flyer posted at a regional police academy. Their average age was 28.26 years ($SD = 7.42$, range: 19–47). Although 16 recruits had served in the military, none had any civilian policing experience.

Materials and Stimuli

Based on a preliminary analysis of the data from Studies 1A and 1B, the experimenter identified three occlusion points for each of the 22 test video scenarios (see Appendix D). For example, in Study 1A the Blow Up scenario was temporally occluded immediately prior to the suspect detonating the improvised explosive device. At that late point in the scenario, all of the experienced participants anticipated the outcome

correctly, as did 71% of the less-experienced participants. In Study 1B, the Blow Up scenario was temporally occluded at an earlier point—when the improvised explosive device had just become visible under the suspect’s shirt. At that relatively early point, only 18.75% of experienced participants and 13.33% of less-experienced participants anticipated the outcome correctly. Based on these data, and assuming that anticipation performance would only improve over the period between the early and late occlusion points, it was reasonable to hypothesize that the point of maximum discriminability would occur during the time period between those two occlusion points. Therefore, the experimenter identified three intermediate occlusion points, each of which corresponded with the onset of a critical cue in the video (see Appendix C). Occlusion points in the remaining scenarios were identified using a similar process. Note that for several scenarios, at least one occlusion point selected for Study 1C corresponded with an occlusion point used in Studies 1A or 1B.

Three temporally-occluded versions of each scenario (i.e., one version for each occlusion point) were created in WMV format using Windows Movie Maker 2.6 (Microsoft); these videos were then converted to SWF (i.e. Flash video) format using Hamster Free Video Converter 2.0 (www.hamstersoft.com). The edited videos were inserted into a web-based experiment using Qualtrics Online Survey Software (www.qualtrics.com). The experiment comprised: (a) an informed consent form, (b) a web-browser check, (c) demographic and law enforcement experience questionnaires, (d) detailed video-based instructions and a single practice trial, (e) test trials, and (f) a debrief and “thank you” note.

Procedure

Experienced participants completed the experiment at their convenience (e.g., at work or at home). Less-experienced participants completed the experiment in a computer lab at the police academy. Participants viewed one *version* of each of the 22 test trials (i.e., one of the three occlusion points for each trial) in a randomized order. As the primary purpose of this study was to determine the optimal occlusion point for each scenario—and not to conduct a within-participant analysis of the effect of occlusion point on performance—the test trials were not blocked by version (i.e., occlusion point number). Instead, the version of a given scenario that was presented to the participant was determined by the software, which was programmed to balance—within each experience group—the number of participants who viewed each version of a scenario. By programming the experiment in this way, version-type was essentially randomized together with scenario-presentation order, and therefore there was no need to counterbalance for order effects associated with version-type. Across the 22 test trials, participants viewed an average of approximately 7.33 scenarios ($SD \approx 2.00$) at each of the three occlusion points.

After viewing the temporally-occluded scenario in each trial, participants typed their anticipated outcome into a text box on the screen and then indicated whether they had seen that scenario previously in their law enforcement training. Participants did not receive feedback on their performance (i.e., they were never told nor shown how the scenario culminated after the occlusion point).

Data Preparation

The data were downloaded from the Qualtrics web site as an SPSS data file, and

were then cleaned and exported to Microsoft Excel for scoring.

DATA ANALYSIS FOR STUDIES 1A–C

Data Screening

If participants indicated that they had seen a scenario before (i.e., during their law enforcement training), their response for that scenario was deleted from the data set.

None of the participants in Studies 1A or B indicated that they had seen any of the scenarios before. Of the 2482 data points collected in Study 1C, 44 (1.78%) were deleted.

Anticipation Scores

The criterion anticipation for each scenario was the final action carried out by the suspect after the occlusion point (e.g., detonating the improvised explosive device, placing the knife on ground and surrendering). Anticipations were scored as “correct” if they matched the criterion.

Analyses Employed

Three types of analyses across the three studies were used to identify scenarios that discriminated between the experience groups, and to identify the optimal occlusion point. These analyses were performed for each scenario separately and provide converging evidence regarding the usefulness of each clip for assessing anticipation performance.

Cumulative probability plots. For each scenario, the procedure used to determine the time point of maximum difference between the experienced and less-experienced participants was based on the 2-sample Kolmogorov-Smirnov test (e.g., Goodman, 1954); a similar procedure has been used previously to identify expert–novice

differences in a card-sorting task (Wolf, Dougherty, & Kortemeyer, 2012).

As a first step toward generating the cumulative probability plots required for this analysis, the log of the odds ratio (i.e., number of correct anticipations divided by the number of incorrect anticipations) for experienced and less-experienced participants was calculated at each occlusion point. In instances where there was zero correct or zero incorrect anticipations, a correction was made by adding 0.5 to both numbers (Agresti, 2002). These log odds ratios (y axis) were plotted against occlusion-point time (x axis), centered around the mean occlusion-point time. A linear trend line was then fitted to each group's data. Using the slope and intercept of each line, the probability of a correct anticipation was calculated across the time period bounded by the earliest and latest occlusion points, in increments of 0.01 s. The cumulative probability for the experienced and less-experienced groups was then plotted against time and the time of maximum difference (in favor of the experienced group) was identified and recorded as the optimal occlusion point.

To ascertain the impact of combining the data from the three studies in one analysis, this process was repeated using only the three occlusion points from Study 1C; the results did not differ from those obtained using the five occlusion points from Studies 1A–C.

2 x 2 Pearson's chi-squared tests of independence. To detect whether the number of correct anticipations generated by the experienced and less-experienced groups were independent of each other, Pearson's chi-squared tests of independence were conducted on the data at each occlusion point. Fisher's exact test was calculated when expected cell frequencies were less than 5.

Pearson's chi-squared goodness-of-fit tests. When chi-squared tests of independence are conducted using small sample sizes—such as those that characterize the number of participants at each occlusion point—the type II-error rate can be inflated (Neyman & Pearson, 1933). For example, the power to detect a difference between proportions-correct of .5 (e.g., less-experienced group) and .7 (e.g., experienced group) with a total sample size of 30 is approximately .20; to detect the same difference with a power of .80, a total sample size of 186 would be required. Furthermore, the power to detect a significant difference between proportions increases as the proportions diverge from 0.50. For example, for a fixed sample size, the power to detect a significant difference between proportions of .50 and .70 is less than the power to detect the difference between proportions of .10 and .30, even though the difference in proportions (i.e., .20) is the same. This is problematic, because in the current study the difference between .50 and .70 is more interesting (i.e., one group performs better than chance) than the difference between .10 and .30 (i.e., both groups performing below chance).

However, using a chi-squared goodness-of-fit test (cf. chi-squared test of independence), researchers can assess differences between observed data and a theoretically-derived expected value. In this analysis, the observed values within each experience group were compared to a distribution in which 50% of participants anticipated the outcome correctly. Fifty percent was selected as the theoretical comparison value as this represents the most conservative comparison value for anticipating at chance levels. Although there is clearly more than one incorrect anticipation that participants could generate, using a lower theoretical comparison value for chance anticipation (e.g., 20%) could result in a low—and potentially uninteresting—

proportion of correct anticipations appearing to be a significant departure from the theoretical expected value (e.g., when greater than 20% but less than 50% of participants anticipate correctly). Using the goodness-of-fit test with the expected value set to 50% of the number of observations, the ideal situation would be the detection of an occlusion point at which the proportion of correct anticipations generated by experienced participants is above, and significantly different from 50%, and the proportion of correct anticipations generated by less-experienced participants is not significantly different from 50%.

Identification of Discriminating Scenarios

Each scenario was initially classified as *discriminating* or *non-discriminating* using the cumulative probability data. Discriminating scenarios were those for which, at any time point(s):

- The experienced group's probability of anticipating the actual outcome was greater than .5, and
- The difference between the experienced and less-experienced groups' probabilities was greater than .1.

Of the 23 scenarios, six were classified as discriminating and were retained for further analysis (see Table 2); the 17 non-discriminating scenarios were not subjected to further analysis.

Using the cumulative probability data for each discriminating scenario, the time point was located that corresponded with the maximum difference between the experienced and less-experienced groups' probabilities. This time point was then compared to the occlusion points used in each study, and the temporally-closest occlusion

point was identified. The occlusion point was classified as *optimal* if:

- The Pearson's chi-squared test of independence was significant at $p < .05$, and/or
- When more than half of the experienced and less-experienced participants anticipated the outcome correctly, the Pearson's chi-squared goodness-of-fit test for the experienced group indicated that the proportion of correct anticipations was significantly greater than 50% and the goodness-of-fit test for the less-experienced group indicated that the proportion of correct anticipations was not significantly different than 50%, and/or
- More than half of the experienced participants, but fewer than half of the less-experienced participants anticipated the outcome correctly.

Using these criteria, all six occlusion points (i.e., one for each discriminating scenario) were classified as optimal. Table 2 displays the optimal occlusion point and results of the classification criteria for the discriminating scenarios.

Table 2
Optimal Occlusion Point for Discriminating Scenarios, with Results of Classification Criteria

Scenario	Optimal occlusion point (s)	χ^2 test of independence ($df=2$) ^d	Group size and percent correct anticipations				χ^2 goodness-of-fit test ($df=1$)			
			Experienced		Less-experienced		Experienced		Less-experienced	
			<i>n</i>	%	<i>n</i>	%	χ^2	<i>p</i>	χ^2	<i>p</i>
Blow up	42.80 ^a	-	11	100	14	57	11.00	<.001	0.29	.593
Domestic punch	28.27 ^c	1.712	14	79	21	57	4.57	.033	0.43	.513
Gun grab	41.60 ^c	-	15	93	22	55	11.27	<.001	0.18	.670
Hey Mr Wilson	66.73 ^b	-	16	81	15	67	6.25	.012	1.67	.197
Subway	31.73 ^c	-	14	71	22	41	2.57	.109	0.73	.394
Suicide waitress	41.20 ^c	-	13	77	22	36	3.77	.052	1.64	.201

^aOcclusion point from Study 1A: Participants interacted with the video scenario and completed option generation tasks. ^bOcclusion point from Study 1B: Participants interacted with the video scenario and generated a single anticipation. ^cOcclusion point from Study 1C: Participants viewed the video scenario passively (i.e., without interacting), and generated a single anticipation. ^dWhen at least one expected cell frequency was less than 5, the χ^2 statistic is omitted and the two-tailed probability for Fisher's exact test is reported.

Effect Size Calculation

For each discriminating scenario, a measure of effect size for anticipation performance was calculated using Cohen's index h for proportions (Cohen, 1988, equation 6.2.2) (see Table 3). Cohen defined small, medium, and large effects as h values of 0.2, 0.5, and 0.8, respectively. Across the six scenarios, the overall sample-weighted mean h was 0.75.

Table 3
Effect Size (Cohen's h) Measures for the Discriminating Scenarios and the Overall Sample-Weighted Effect Size

Scenario	Cohen's h	Total n	Weighted h
Blow up	1.43	25	35.76
Domestic punch	0.48	35	16.74
Gun grab	0.94	37	34.60
Hey Mr Wilson	0.32	31	9.98
Subway	0.61	36	22.12
Suicide waitress	0.85	35	29.90
		$\Sigma = 199$	$\Sigma = 149.09$
Overall sample-weighted h :	0.75		

SUMMARY OF STUDIES 1A–C

The goals of Study 1 were to identify video-based law enforcement scenarios that could discriminate between experienced and less-experienced police officers' ability to perceptually anticipate the outcome, and determine the optimal occlusion point that maximized the skill-based difference in anticipation. Three studies, all employing a temporal-occlusion paradigm, were conducted to achieve these goals. Across the three

studies, an iterative approach was taken to the selection of occlusion points. In Study 1A, the experimenter selected the occlusion point. In Study 1B, a cognitive task analysis, using experienced police trainers, was conducted to identify the criterion occlusion point. Then, based on the data from Studies 1A and 1B, Study 1C employed three different occlusion points in a between-subjects design.

The data from the three studies were analyzed collectively using converging methods. Of the 23 scenarios that were examined, the analysis revealed six scenarios to be discriminating of experience level. For the six discriminating scenarios, an optimal occlusion point—which maximized the difference between experienced and less-experienced participants in terms of their anticipation performance—was determined. The discriminating scenarios—occluded at their respective optimal occlusion points—served as the experimental stimuli in Study 2.

Note that a secondary goal of Study 1A was to assess the option-generation paradigm for use in Study 2. In general, the option-generation paradigm was effective in eliciting the data required for testing hypotheses related to option generation during the assessment and response phases of decision making. Several minor modifications that could improve the paradigm were identified (e.g., forcing participants to assign a unique likelihood rating [e.g., 51% and 49%, rather than 50% and 50%] to each option in a set); the revised paradigm was employed in Study 2.

Finally, the methods and analyses employed in Study 1 provide an example of a procedure that can be used to determine occlusion points for temporal-occlusion studies, especially in domains where the ultimate outcome of a given action is not presaged solely by physically deterministic events (e.g., biomechanical properties of action), but instead

could also include probabilistic information (e.g., pulling a gun, versus a wallet, out of a pocket).

CHAPTER 6

STUDY 2

Recall that the main goal of this research is to further our understanding of the relationship between the situation assessment and response phases of decision making, and to investigate the extent to which this relationship is affected by skill level. An additional goal was to gather information that can be used, in the future, to design training based on the anticipation skills, decision strategies, and processes of experienced police officers.

Study 1 succeeded in identifying six escalating scenarios—and optimal occlusion points—that discriminated between experienced and less-experienced officers, in terms of their ability to anticipate the outcome of the scenario. Using these scenarios, Study 2 was designed to test specific hypotheses about option generation during the situation assessment and responses phases of decision making. The hypotheses, which are derived from LTWM theory and the RPD/TTF model, focus on the number and type (i.e., task-relevant versus task-irrelevant) of options generated, the serial position of the “selected” (i.e., final) option, and the relationship between the number of options generated and decision quality. To collect data (i.e., options, ratings of options) that could be used to test these hypotheses, Study 2 employed a prediction and option-generation paradigm (similar to that described for Study 1A), in which participants viewed temporally occluded scenarios, and then generated assessment and response options.

Although it was expected that the option-generation and ratings data would provide a detailed insight into officers’ thinking at the point of occlusion, these data alone do not provide a comprehensive understanding of information processing during these

dynamic scenarios. For example, an examination of the options generated at the occlusion point does not—in and of itself—provide information about other, additional options that participants may have generated earlier in the scenario but then discounted prior to the occlusion point. Similarly, it would be interesting—especially for the future design of training—to gain an understanding of what led participants to generate the options they did. Therefore, to supplement the option generation and rating data collected at the occlusion point, additional information about participants' thinking was elicited via two cognitive task analysis techniques: retrospective verbal reports of thinking, and stimulated recall using probe questions. The data elicited via these techniques were expected to validate the option generation and ratings data, and provide information about cues heeded, and inferences and options generated, prior to the occlusion point.

Hypotheses

Hypotheses 1 and 2 relate to outcome measures of performance.

Hypothesis 1: Based on the review of studies of perceptual anticipation, it was expected that experienced participants would anticipate the culminating event more frequently than less-experienced participants.

Hypothesis 2: Based on the review of response-option generation and selection studies, it was expected that experienced participants would select the best response option more frequently than less-experienced participants.

Hypotheses 3–8 relate to the option-generation process.

Hypothesis 3. Based on previous studies that employed an option-generation paradigm (e.g., Johnson & Raab, 2003; Klein & Peio, 1989; Klein et al., 1995; Raab & Johnson,

2007; Ward et al., 2012), it is expected that in general, participants would generate few (i.e., less than five) assessment and response options.

Hypothesis 4: Based on LTWM theory, Ward et al. (2011) proposed that possessing a detailed representation of the situation (i.e., generating more assessment options) should facilitate the generation of a single, good response option. Although these authors did not find support for this hypothesis, methodological differences between this and other option generation studies could explain their results. Accordingly, it was still expected that in the current task environment the total number of assessment options would be negatively correlated with the total number of response options.

Hypothesis 5: Recall that Ward and Williams (2003) found that skilled youth soccer players generated more task-relevant options than task-irrelevant options, and that compared with less-skilled players, they generated fewer task-irrelevant options. To explain this finding, Ward & Williams proposed that over time, skilled players develop extensive, detailed, and domain-specific knowledge representations, which they can access via retrieval cues in LTWM. In dynamic situations, skilled players create up-to-date situation models by integrating information from the environment with their knowledge representations. The situational model aids in identifying meaningful patterns of cues in the environment, while disregarding irrelevant information. Less-skilled players generated more irrelevant information because they had not developed rich knowledge representations, nor did they possess the LTWM skills needed to integrate such knowledge with the information available in the environment. Based on this explanation, Ward et al. (2011) hypothesized that skilled police officers would generate more task-relevant assessment options than task-irrelevant assessment options, and that

they would generate less task-irrelevant assessment options than less-skilled officers. The data, which was extracted from verbal reports, supported this hypothesis. Interestingly, although this hypothesis relates specifically to options generated during assessment of a situation, the same pattern of results was also found for response options—despite the more general claim by proponents of TTF and RPD that experts would generate fewer *total* response options during decision making (e.g., Johnson & Raab, 2003; Klein & Peio, 1989). Based on Ward & Williams, and Ward et al., it was expected that experienced police officers will generate more task-relevant options than task-irrelevant options, and that they will generate fewer task-irrelevant assessment options than less-experienced officers during both assessment and responses phases of decision making.

Hypothesis 6: Recall that Ward and Williams (2003) and Ward et al. (2012) found that skilled soccer players were better at prioritizing—by threat level—their task-relevant assessment options. This ability was attributed to skilled players' superior domain-specific knowledge representations and LTWM skills. Based on this finding, it is expected that compared to less-experienced officers, experienced police officers will be better at prioritizing their task-relevant assessment options by threat (i.e., danger posed to the officer and innocent bystanders).

Hypothesis 7: Based on LTWM theory, Ward et al. (2011, 2012) hypothesized that the ability to generate more of the task-relevant assessment options—when they are available in the environment—would result in better anticipation performance. In other words, having a detailed and up-to-date situation model would allow participants to accurately anticipate the actual outcome. Conversely, Ward et al. hypothesized that generating more task-irrelevant assessment options would result in poorer anticipation performance.

Indeed, the data supported these claims. Therefore, it is expected that the number of task-relevant assessment options generated will be positively correlated with the number of times participants correctly anticipate the situational outcome. It is also expected that the number of task-irrelevant assessment options will be negatively correlated with the number of times participants correctly anticipate the situational outcome. Additionally, it is expected that the total number of options (i.e., task-relevant plus task-irrelevant) will not be correlated with anticipation performance, since the underlying positive and negative relationships will nullify each other, resulting in no significant relationship.

Hypothesis 8: According to TTF, the total number of response options should be negatively correlated with decision quality. Although Johnson and Raab (2003, Raab & Johnson, 2007) found support for this notion, Ward et al. (2011) suggested that LTWM theory could provide a more nuanced understanding of the relationship between the number of response options generated, and the quality of the final decision. Accordingly, Ward et al. found that the number of task-irrelevant response options generated was negatively correlated with decision quality, but that the number of task-relevant response options was positively correlated with decision quality. Based on the opposing relationships observed during assessment and response, Ward et al. also found that the total number of response options was not significantly correlated with decision quality. Thus, if a TTF-like mechanism underlies response-option generation, it is expected that the total number of response options will be negatively correlated with the number of times participants select the best response option. If, on the other hand, a LTWM-type mechanism underlies response-option generation, it is expected that the number of task-

irrelevant and task-relevant options will be negative and positively correlated, respectively, with the number of times participants select the best response option.

Method

Power Analysis

To determine the sample size required for Study 2, an a priori power analysis was conducted using G*Power 3.1.5 software (Faul, Erdfelder, Lang, & Buchner, 2007), with α set at .05 and β set at .80. The option to input effect size as Cohen's h was selected, and the sample-weighted mean proportion of correct anticipations for the less-experienced group in Study 1 (i.e., .51) was entered as the lower expected proportion. G*Power estimated the sample size required to detect an effect size of $h = 0.75$ (i.e., the overall, sample-weighted effect size from Study 1) under these conditions to be 46 (i.e., 23 per experience group) (see Appendix F).

Participants

Experienced participants ($n = 23$, 1 female, 22 males) were law enforcement officers from a metropolitan police department who had accrued at least 5 years' law enforcement experience ($M = 15.29$, $SD = 4.62$) and were, at the time of testing, serving in an active role (e.g., street patrol, special response team member, undercover officer). These officers had completed one week of refresher training per year of service, as mandated by their department. Additionally, eight of the officers had served in the military; four of those served in combat-related roles. The median age of the experienced officers was 40 years (range: 33–59). Apart from the 23 officers who were included in the sample, four other experienced officers participated in the experiment but were excluded from the experienced group because they had either: (a) accrued less than 5 years' law

enforcement experience ($n = 2$), (b) retired from law enforcement ($n = 1$), or (c) recently returned to working street patrol after working in administrative roles for more than 10 years ($n = 1$).

Less-experienced participants ($n = 26$, 6 females, 20 males) were law enforcement trainees who were in the process of completing ($n = 20$), or had recently completed ($n = 6$), their initial law enforcement training. Their median age was 22 years (range: 21–40). The trainees had not yet begun working as fully-fledged police officers, but had spent at least 176 hours either observing an experienced police officer on duty ($n = 20$) or carrying out police duties whilst under the supervision of an experienced police officer ($n = 6$). Three other trainees completed the experiment, but were excluded from the less-experienced group because they had prior law enforcement, military, and/or security experience.

Materials and Stimuli

Ten interactive, video-based law enforcement scenarios (IES Interactive Training: Ann Arbor, MI) served as the basis for the temporally-occluded experimental stimuli. One escalating scenario was used as the practice trial; this was the same practice trial used in Study 1. Six *test* scenarios were the escalating scenarios from Study 1 that discriminated between skill groups on anticipation performance. The test scenarios were temporally occluded at the optimal occlusion point determined by the analysis of the Study 1 data. Three de-escalating scenarios served as catch trials. The catch trials, and appropriate occlusion points, were selected by a subject-matter expert (SME) in police use-of-force who identified scenarios from Study 1 in which a responding officer would be justified in placing their hand on their holstered firearm while questioning the

suspect(s), but would not be justified in using any force (i.e., drawing a firearm/shooting, drawing a baton/striking, drawing chemical spray/spraying, or employing empty-hand strikes/physical restraint techniques). The subjective judgments of the SME were consistent with the use-of-force continuum that forms the basis of current law enforcement practice and training for these kinds of decision-making scenarios.

Each scenario was digitally edited to end at the occlusion point (i.e., replaced by black screen), thus creating the video stimuli. An audible tone was inserted immediately after the occlusion point; this served to prompt participants to begin the subsequent option generation tasks. As in Study 1, each video stimulus began with an audio-only “radio message”, describing the situation to which the officer was being dispatched. Each video stimulus was inserted into a Microsoft PowerPoint 2010 presentation, which included prompts for the subsequent assessment and response option generation tasks.

Option Generation Tasks

Assessment option-generation task. In this task, participants were required to generate assessment options—events that could occur in the environment (e.g., courses of actions that could be taken by the perpetrator or others in the scenario) at the point of occlusion. To elicit assessment options, participants responded verbally to the following textual prompt, which was displayed on the screen: “What could happen next on the screen in the next few seconds?” The prompt was displayed together with a 10-second countdown timer and a single text box. Participants were instructed to verbalize their first option as quickly as possible, and that their response time would be measured, starting from the occlusion-point tone. The experimenter transcribed the option into the text box. This process was repeated using a new slide/timer for each additional option, until the

participant indicated that they had verbalized all of the assessment options that were heeded at the point of occlusion. To minimize the extent to which participants deliberated about options prior to verbalization, the experimenter emphasized—during the practice trial—that participants should: (a) list, rapidly, only those assessment/response options that they were considering at the point of occlusion; (b) verbalize all of the assessment/response options under consideration, even if they believe a specific option(s) to have a low likelihood; and (c) refrain from engaging in reflective option generation (e.g., reporting options that were generated only after reflecting on the scenario after it had been occluded, generating options solely for the sake of increasing the number of options generated, etc.).

Then, to facilitate the subsequent rating tasks, the complete set of transcribed options was displayed on a single screen; a small text box next to each option was used to record the ratings. Participants first rated the likelihood that each of their generated options would actually happen after the occlusion point, using a scale that ranged from 0 (*not at all likely*) to 100 (*very likely*). To avoid potential subadditivity or superadditivity of the likelihood ratings, and to facilitate a subsequent analysis of entropy (i.e., information value), the likelihood ratings were required to sum to 100. To assist participants in this task, a “remainder” box at the bottom of the screen displayed the number of to-be-allocated likelihood points; this number updated automatically each time the experimenter transcribed a likelihood rating next to its corresponding option.

Next, using a new screen with blank rating boxes, participants rated how threatening each option was to their own personal safety (i.e., if they were actually dealing with the incident) using a scale from 0 (*not at all threatening*) to 100 (*very*

threatening). As participants could have generated more than one option that they rate as very low/high threat, the threat rating for each option was made independently (i.e., without regard to the sum of the individual threat ratings).

To facilitate identification of the most-likely and most-threatening options during data analysis, participants were instructed to provide unique likelihood ratings and unique threat ratings. For example, in the case where two options were generated, participants were required to assign likelihood ratings that differed by at least one point (e.g., 51 and 49, instead of 50 and 50).

The set of assessment options displayed on the likelihood- and threat-rating screens were presented in a shuffled order, and not in their order of generation. This step was implemented to reduce the possibility that participants would simply rate options according to their serial position on-screen (e.g., assign a high likelihood to the first option generated, and a lower likelihood to the next option, etc.). Note, however, that the serial order of option generation was recorded for later analysis.

Response option generation task. Participants completed a similarly structured response option generation task by answering the question, “How could you respond in the next few seconds?” After verbalizing their response options, participants first rated their likelihood of pursuing each option, employing the same likelihood rating scale and procedure described for the assessment option generation task. Next, they rated the quality of each option using a scale that ranged from 0 (*not at all good for my own immediate safety and survival*) to 100 (*very good for my own immediate safety and survival*). Similar to the threat ratings for assessment options, the option-quality ratings

were made independently for each response option. Participants were required to assign unique likelihood ratings and unique quality ratings to their response options.

Post-Option-Generation Cognitive Task Analysis

As the occlusion points for the test trials occurred relatively late in the scenarios, it was possible that participants—especially those in the experienced group—would find the outcome of the situation to appear obvious. In such cases, participants may verbalize only one assessment option and one corresponding response option. However, prior to the occlusion point, those participants may have generated (and then discarded) other assessment/response options. Therefore, to provide a possible explanation for cases in which a single assessment/response option is generated at the occlusion point, and to gather information that could inform the design of decision-making training in the future, two cognitive task analysis (CTA) techniques (Crandall, Klein, & Hoffman, 2006)—elicitation of retrospective verbal reports (Ericsson & Simon, 1993) followed by stimulated recall of in-event option generation—were used to supplement the option-generation and ratings data.

Use of the retrospective verbal report technique was intended to provide an objective record of participants' thinking while they watched the video stimulus during the test trials. The aim of the subsequent stimulated recall was to elicit additional information about (a) cues heeded, (b) associative inferences generated, (c) specific assessment- and response-options generated that were not reported during the experimental tasks or in the retrospective report, and (d) prior knowledge and/or experience that might have influenced the participant. Due to the time-consuming nature of these techniques, they were administered to each participant on two of the six test

trials. Across all participants/trials, approximately eight participants from each experience level completed the post-option-generation cognitive task analysis for each test trial; the number of participants (i.e., ≈ 8 at each experience level) varied slightly between trials due to a difference in the number of experienced and less-experienced participants who completed the experiment, and the subsequent exclusion—post data-collection—of several participants from each group.

Retrospective verbal reports. Due to the limited time that was available for data collection, abbreviated verbal reporting instructions were administered that were consistent with Ericsson and Simon (1993, pp. 375–379). Specifically, the experimenter instructed participants to (i) recount each thought they recalled having, in the order they had them, from the beginning of the scenario until the occlusion point, starting with the first thought they could specifically remember having; (ii) report only those thoughts that they could specifically recall thinking while the video was playing; and (iii) refrain from explaining, describing, or reflecting on their thinking after the fact.

Stimulated recall of in-event option generation. Following elicitation of the retrospective verbal report, the stimulated recall procedure was used to reveal otherwise unreported instances of assessment- and/or response-option generation, and to stimulate recall of cues heeded, inferences made, and prior knowledge used during the scenario, but prior to the occlusion point. The video acted as a timeline of events that served to cue participants about what they were thinking at a specific point. Prior to replaying the video stimulus, the experimenter instructed participants to say “Stop” when they identified a point in the video at which they recalled noticing something in the situation that they attended specifically and/or a specific action or course of action they contemplated

pursuing. When the participant indicated such a point, the experimenter paused the video and administered probes that were adapted from those described by Crandall, Klein, and Hoffman (2006, p. 79).

Specifically, to identify specific cues that the participant heeded, the experimenter asked, “What were you paying attention to at this point?” To elicit evaluations and associative inferences that the participant generated, the experimenter asked, “You said that you were paying attention to X. What did that mean to you at this point? What was your understanding of the situation at this point?” Given that level 1 and 2 verbal reports had already been elicited via the retrospective reports, prompting participants to provide level 3-type reports (e.g., reflective, interpretative, etc.) during this subsequent CTA task did not contaminate their original memory for the events. Although it might not be scientifically defensible to use the latter data to generate causal explanations for participants’ behavior, they have much value in generating potential hypotheses about the underlying cognitive process and/or representation that could be tested in future research (see de Groot, 1965).

To elicit the participant’s anticipations about the assessment outcome, the experimenter asked, “At this point, what did you think would happen in the next few seconds?” To elicit information about response options considered, the experimenter asked, “What course(s) of action were you considering at this point? What led you to consider that option, or reject other options? Was there a rule that you followed that led you to consider that option?” To identify prior knowledge and/or experience used by the participant to guide their assessment/decision making, the experimenter asked, “What prior knowledge or personal experience influenced your thinking at this point?” After the

participant answered the probes, the experimenter continued the replay-pause-probe sequence until the end of the video stimulus (i.e., pausing at multiple points, as directed by the participant).

Procedure

After providing their written informed consent, participants completed a biographical questionnaire, including questions about their law enforcement training, qualifications, and employment, as well as the number of times they had participated in law enforcement training using video-based simulations. Participants sat 2 m from a large projector screen, and wore a lapel microphone (Sony WCS-999) that was connected, wirelessly, to a video camera (Canon VIXIA HF M300). The camera recorded the participant, the experimenter's and the participant's verbal communication, as well as the projected video stimuli.

The experimenter instructed participants to observe silently (i.e., not interact with) each video stimulus from the perspective of an officer responding to the incident. Participants then viewed the practice trial; after the occlusion point, the experimenter introduced the assessment and response option-generation tasks. Participants then completed the assessment and response option-generation tasks for the practice trial. After the practice trial, participants completed the six test trials and three catch trials in a randomized order. To control for task order effects, the order in which participants completed the assessment and response option-generation tasks was counterbalanced within each experience group. Each participant followed the same task order (i.e., either assessment/response or response/assessment) across all trials, including the practice trial. After completing the option-generation tasks in a given trial, participants indicated

whether they had seen that specific video scenario previously in their law enforcement training. Finally, on selected trials, participants completed the post-option-generation, cognitive task analysis.

Participants did not receive feedback on their performance (i.e., they were never told nor shown how the scenario culminated after the occlusion point). After each trial, the transcribed data from the option-generation tasks was automatically written to a text file, and saved for later extraction and analysis. Participants were tested individually; each testing session lasted approximately 120 minutes. At the conclusion of each testing session, the experimenter debriefed the participant and thanked them for their participation. Participants were requested to refrain from discussing the scenarios—and the options they generated—with other potential participants.

Data Analysis

None of the participants indicated that they had seen any of the video scenarios previously in their law enforcement training; therefore, all data were retained for analysis.

Signal detection analysis. As a preliminary step, a signal detection analysis was employed to determine whether participants were biased toward using lethal force (i.e., drawing and/or shooting their gun) and could discriminate between trials. A SME classified all scenarios as either those in which a shooting/draw gun response was *justified* ($n = 4$ test trials) or *unjustified* ($n = 5$; 2 test trials, plus the three catch trials). For each trial, participant's most likely response option was coded as either *drew gun/shot* (1), or *did not draw gun/shoot* (0). After pooling the data within each experience level, the proportions of hits, misses, correct rejections, and false alarms were calculated. This approach (i.e., pooling the data, rather than calculating proportions for each participant),

was suggested by Macmillan and Creelman (2005) as appropriate for situations in which there were few data points—here there were nine—for each participant. Then, using the formulae described by Stanislaw and Todorov (1999), parametric d' and non-parametric A' were calculated as measures of sensitivity; parametric β , c (the deviation of β from an ideal observer, measured in standard deviation units), and non-parametric B'' were calculated as measures of response bias.

The parametric measures d' and β were interpreted using the guidelines described by Proctor and Van Zandt (2008): A d' of 0.0 is characteristic of performance at the chance level (i.e., completely insensitive), and a d' of 2.33 represents near-perfect sensitivity. A β of 1.0 represents an unbiased response, a β that is greater than 1.0 represents a conservative response, and a β that is less than 1.0 represents a liberal response. The additional, parametric bias measure c was interpreted using the guideline described by Stanislaw and Todorov (1999): a c value of 0 represents an unbiased response, a positive c value represents a bias toward a *no* (i.e., conservative) response, and a negative c value represents a bias toward a *yes* (i.e., liberal) response.

The non-parametric measures were interpreted using the guidelines described by Stanislaw and Todorov (1999): An A' of .5 is characteristic of performance at the chance level (i.e., completely insensitive), and an A' of 1 represents near-perfect sensitivity. A B'' value of 0 represents an unbiased response, a B'' value of -1 represents an extremely liberal response, and a B'' value of 1 represents an extremely conservative response.

Dependent variables to test the experimental hypotheses. The tests of the experimental hypotheses focused on the data from the six test trials. The following dependent variables were defined.

Total number of assessment/response options. This variable was defined as the total number of assessment/response options generated across the test trials.

Type of information: Number of task-relevant and -irrelevant options. To calculate the frequency of task-relevant and -irrelevant options generated across the test trials, and to assess the participants' ability to prioritize the task-relevant assessment options, authoritative ratings of each assessment/response option were elicited from a SME (i.e., a civilian expert who had extensive experience investigating police use-of-force incidents). The SME:

1. Viewed each video stimulus in the same manner as the participants (i.e., until the point of occlusion).
2. Subsequently viewed the culmination of the scenario (i.e., the events after the point of occlusion).
3. Reviewed any part of the scenario they wished, using the rewind, pause, and slow motion functions.
4. Completed the assessment/response option-generation and rating tasks based on the information available up until the occlusion point. Steps 2 and 3 were implemented so that the SME could carry out this step with a comprehensive, contextualized understanding of the events observed up until the occlusion point.
5. For each scenario, reviewed a compiled list of all of the unique assessment/response options generated by participants. Then, using their own options/ratings as a guide, the SME assigned likelihood/threat ratings to each assessment option on the list, and likelihood/goodness ratings to each response option. Effectively, this process allowed the SME to identify any additional assessment options that he had not generated

himself, but considered, nonetheless, to be likely and threatening to some degree. Similarly, the SME could identify additional response options that he considered likely and good courses of action to pursue.

The SME completed these steps after the data collection phase of Study 1A. The SME viewed the occlusion points from Study 1A, and reviewed compiled lists of assessment/response options generated by the participants in Study 1A. Note that the SME provided ratings for four of the six test trials used in the current study; two scenarios—Hey Mr Wilson and Suicide Waitress—were included as stimuli only after the conclusion of Study 1A.

Task-relevant assessment options were defined as the options generated by the SME during step 4, and any options from the compiled list that the SME judged as possible (i.e., likelihood >0%) and potentially threatening (i.e., threat >0%). Task-irrelevant assessment options were options from the compiled list that the SME rated as not at all likely and/or not at all threatening (i.e., 0% likely and/or 0% threatening). Task-relevant and irrelevant response options were defined by applying the same criteria to the likelihood/goodness ratings.

To determine a measure of inter-rater reliability, the experimenter subsequently reviewed the compiled lists of assessment/response options that were presented to the SME in step 5. Based on the occlusion points used in the current study, the experimenter indicated whether each option was task-relevant or task-irrelevant. For the four test trials rated by the SME and the experimenter, inter-rater reliability (Cohen's kappa) was .72 and .77 for assessment and response options, respectively; according to Landis and Koch (1977), these values signify substantial agreement. Based on the experimenter's

familiarity with the occlusion points used in the current study, and the knowledge accumulated as a result of conducting CTAs of these scenarios with several other SMEs (see Study 1B), the experimenter's ratings of task relevance were used as the criterion ratings. The experimenter also determined the criterion task-relevance ratings for the two test trials not included in Study 1A.

Prioritization of task-relevant assessment options. Participants' ability to prioritize the task-relevant assessment options in each test trial was measured based on the SME's threat rating for each task-relevant option. Although SME and participant threat ratings were elicited using a continuous scale, the threat rating data were first converted to threat ranks (i.e., most threatening, second-most threatening, etc.). This step was taken to reduce the impact of subjective differences—between participants, and between the participants and the SME—in the absolute threat ratings. The threat ranks encapsulated the pertinent information: the relative differences in perceived threat amongst the options generated.

Based on the method described by Ward et al. (2012), a weighted threat prioritization score was calculated for each participant in each trial. If participants generated the SME's most-threatening option as their first assessment option, and also ranked it as their most-threatening option, they received five points. If they generated the SME's second-most threatening option as their second assessment option, and then ranked it as their second-most threatening option, they received four points, etc. For each deviation from the SME's rank—in terms of serial position of generation and threat ranking—participants were penalized one point. If participants failed to generate one of the SME's task-relevant options, they received zero points for that option. For each trial,

the number of points awarded for each of the SME's task-relevant options was summed, and the sum was then divided by the number of possible points, to yield a prioritization score with a minimum value of 0 and a maximum value of 1.

Predict the first assessment option as the outcome. This variable was defined as the number of times that the first assessment option generated was rated as the option most likely to happen, irrespective of whether this corresponded with the culminating event or the most threatening option.

Predict the culminating event. This variable was defined as the number of times that the culminating event was the option that was rated as most likely to occur next, irrespective of the serial position of option generation.

Predict the culminating event first. This variable was defined as the number of times that the culminating event was generated as the first option and predicted to be the most likely to happen next.

Take the first response option. This variable was defined as the number of times that the first response option generated was rated as the option most likely to be pursued, irrespective of how the SME rated that option and of how they, the participant, rated the quality of that option.

Take the best response option. This variable was defined as the number of times that the best option identified via the SME ratings was taken by the participant (i.e., rated as the likely one they will pursue), regardless of serial position and their ratings of quality.

Take the best response option first. This variable was defined as the number of times that the best option identified via the SME ratings was taken by the participant (i.e.,

rated as the likely one they will pursue) and generated as their first option, regardless of their ratings of quality.

Tests of the experimental hypotheses. To test the hypotheses, the following analyses were conducted. Note that as a preliminary step, all analyses were conducted with task order (i.e., assessment task first or response task first) as an additional between-subjects factor. Across all analyses, there was no significant main effect of task order, nor any significant interactions that included task order. As there was no theoretical reason to include task order as a factor of interest, it was dropped from all analyses and will not be referred to further. *P* was set at .05.

Hypothesis 1: Effect of experience on the ability to predict the actual culminating event. Independent samples *t*-tests were used to assess whether experience affected participants' ability to predict the actual culminating event, and to predict the culminating event first.

Hypothesis 2: Effect of experience on the ability to take the best response option. Independent samples *t*-tests were used to assess whether experience affected participants' ability to take the best response option, and to take the best response option first.

Hypothesis 3: Few assessment/response options generated. The total number of assessment/response options was compared to those found by previous researchers (e.g., Johnson & Raab, 2003; Klein & Peio, 1989; Klein et al., 1995; Raab & Johnson, 2007; Ward et al., 2012). No formal statistical tests were conducted for this hypothesis.

Hypothesis 4: Negative relationship between the total number of assessment and response options. A correlational analysis was used to assess the relationship

between the total number of assessment options and the total number of response options.

Hypothesis 5: Differences in the numbers of task-relevant and task-irrelevant options generated. A 2×2 multivariate analysis of variance (MANOVA) was used to assess the effects of experience (between-subjects factor) and type of information (within-subjects factor) on the numbers of assessment and response options generated. Significant multivariate effects were followed up using separate univariate analyses of variance (ANOVAs) for assessment/response options separately.

Hypothesis 6: Threat prioritization of task-relevant assessment options. An independent samples *t*-test was used to assess whether experience affected participants' ability to rank their task-relevant assessment options by threat.

Hypothesis 7: Relationship between the number of task-relevant/irrelevant assessment options and the ability to predict the culminating event. Note that for a given scenario, the culminating event was, by definition, one of the task-relevant assessment options. For this analysis, counting the culminating event as a task-relevant option could result in a spurious positive correlation between the number of task-relevant options and the number of times participants predicted the culminating event. Therefore, prior to the analysis the number of task-relevant options was adjusted as follows: if participants predicted the culminating event in a given trial, their number of task-relevant options was reduced by one.

A multiple linear regression was conducted to determine whether the numbers of task-relevant and -irrelevant assessment options were positively and negatively correlated with the ability to predict the culminating event, respectively. To check whether participants' experience accounted for any additional variance—over and above that

explained by the numbers of task-relevant and task-irrelevant options—experience level was added to the model in a second block. A separate linear regression was used to assess the relationship between the total number of assessment options and the ability to predict the culminating event; experience level was added to the model in a second block.

Hypothesis 8: Relationship between the number of response options and the ability to take the best response option. For a similar reason to that described for Hypothesis 7, the number of task-relevant response options was reduced by one when participants took the best response option.

A linear regression was used to assess the relationship between the total number of response options and the ability to take the best response option; experience level was added to the model in a second block. A separate linear regression was conducted to assess the relationship between the numbers of task-relevant/irrelevant response options and the ability to take the best response option. To check whether participants' experience accounted for any additional variance—over and above that explained by the numbers of task-relevant and task-irrelevant options—experience level was added to the model in a second block.

Supplementary, exploratory analyses on the option generation/rating data. In addition to the tests of the experimental hypotheses, two exploratory analyses were conducted on the option generation/rating data for the six test trials.

Take the first option. The number of times that participants rated their first assessment/responses option as most likely (i.e., the participant chose the first option as the anticipated actual outcome or best decision) was counted, and expressed as a

proportion of the number of test trials. The values were compared, descriptively, with those reported by Johnson and Raab (2003).

Entropy. Shannon (1948) introduced the concept of entropy, H , as a measure of information uncertainty, where:

$$H = - \sum_{i=1}^n P(x_i) \log_b(P(x_i)) \quad (1)$$

Based on equation 1, information uncertainty generally increases as additional options are generated, and as the likelihood ratings become more evenly distributed amongst the generated options. For example, in the current study, when a participant generated a single assessment option, they had to rate it as 100% likely to happen next. Such a case signifies a total lack of uncertainty about the culminating event, and H is equal to zero. If a participant generated two options and rated one option as much more likely than the other—for example, 90% versus 10%—they display greater uncertainty compared with the single-option case, and H is equal to 0.47. Consider a more extreme case of uncertainty, in which a participant generates two options and rates one 51% likely and the other 49% likely. In this situation there is near-total uncertainty about the culminating event, and H is approximately equal to 1.

Entropy scores were calculated for each test trial using participant's assessment/response likelihood ratings. The effect of experience on participants' mean entropy scores during assessment/response were assessed using independent samples t -tests.

CTA data. Recall that for each test trial, CTA data was collected from approximately 16 participants (i.e., eight experienced, eight less-experienced). The

retrospective verbal reports, and the reports elicited via stimulated recall, were transcribed from audio recordings. Subsequently, each participant's reports were segmented using natural speech and other syntactical markers.

Retrospective verbal report data. The number of words in each participant's report was counted, and an independent samples *t*-test was used to determine whether experience level affected the number of words verbalized. Similarly, the number of statements (i.e. segments) in each participant's report was counted, and an independent samples *t*-test was used to determine whether experience level affected the number of statements.

Statements were coded inductively as one of five thought types (i.e., monitored event, inference, evaluation, prediction, action) using procedures outlined in Ericsson (1975, cited in Ericsson & Simon, 1993). These five thought types were a subset of those employed by Ward et al. (2011) (see Appendix G). To determine whether experience level affected the number of times each thought type was reported, an independent samples *t*-test was conducted for each thought type; a Mann-Whitney test was employed when thought-type data were non-normally distributed.

Data elicited via stimulated recall. Based on the probe questions employed during the stimulated recall procedure, the segmented data were coded as one of five thought types (i.e., cue heeded, evaluation/inference made, prediction/response option generated, knowledge used). To facilitate a fine-grained analysis of the data based on events within a scenario, the test trials were scrutinized by the experimenter, who identified 2–3 distinct *phases* in each scenario. The onset of a new phase represented a change in environment (e.g., from being outside a house to entering a house) or an

observable change in a suspect's behavior (e.g., from sitting, silently, on the ground to standing up and being verbally aggressive toward the police). Using a timeline of events in each scenario as a reference, each statement was coded by scenario phase (i.e., first phase, second phase etc.).

Quantitative analysis. Although a 2 (experience level) \times 3 (scenario phase) \times 4 (thought type) mixed ANOVA was planned to analyze these data, a preliminary exploration revealed the existence of non-normal distributions and unequal variances in a substantial number of cells. In light of this, the median values for each thought type were tabulated, and the data are interpreted descriptively.

Qualitative analysis. Recall that the purpose of the stimulated recall procedure was to elicit additional information that might be useful in explaining experience-based differences in the option-generation and rating data. Therefore, each participant's data were reviewed and summarized in an attempt to capture the gist of their thinking. This involved highlighting and linking statements that appeared to explain and/or justify the participant's response option(s).

Then, using the option-generation and rating data for each test trial, the participants in each experience group who completed the stimulated recall procedure were categorized according to whether they predicted the culminating event and whether they took the best response option. This yielded eight possible categories—two categories in each of four cells—as depicted in Table 4.

Table 4
Performance Categories as Functions of Experience Level, Ability to Predict the Culminating Event, and Ability to Take the Best Response Option

Predicted the culminating event	Took the best response option	
	Yes	No
Yes	a. Experienced b. Less-experienced	c. Experienced d. Less-experienced
No	e. Experienced f. Less-experienced	g. Experienced h. Less-experienced

Within each category, the summaries of participants' explanations were compared in order to identify any common themes. Additionally, most of the cells contained some experienced, and some less-experienced, participants; essentially, for a given scenario, there were some experienced participants who predicted/responded incorrectly, and some novices who predicted/responded correctly. This general phenomenon—of some experienced performers acting like less-experienced performers, and vice versa—is not uncommon. For example, across numerous studies of skilled anticipation and decision making in sport, there have been few, if any, instances of experts outperforming novices on every trial (e.g., Williams, Ford, Eccles, & Ward, 2011). This is particularly the case when experienced performers' abilities do not closely match the task demands, and when mere domain experience (i.e., in the absence of deliberate practice) does not lead to improved performance (Camerer & Johnson, 1991; Ericsson & Lehmann, 1996; Shanteau & Stewart, 1992), which may be the case in this study. Therefore, when cells contained experienced and less-experienced participants, the summaries of both groups were contrasted to seek evidence of underlying, experience-based differences. Finally, the

theme(s) in each category/cell were compared with those in the other categories/cells, and any differences were identified. Observed differences were then described qualitatively.

Results

Signal Detection Analysis

The proportions of hits, misses, false alarms, and correct rejections for experienced and less-experienced groups are described in Table 5; there were very few false alarms.

Table 5
Proportion of Times that Participants Drew/Did Not Draw their Gun by Trial Type and Experience Group

Trial Type	Experienced Group		Less-experienced Group	
	Drew gun	Did not draw gun	Drew gun	Did not draw gun
Draw gun: Justified	.42 Hit	.58 Miss	.39 Hit	.61 Miss
Draw gun: Unjustified	.02 False alarm	.98 Correct rejection	.01 False alarm	.99 Correct rejection

The measures of sensitivity and bias calculated from these data are reported in Table 6. The signal detection analyses revealed that experienced and less-experienced participants were very sensitive to differences in trial type: they were good at discriminating between trials in which a “draw gun” response was justified, versus unjustified. Although the d' values indicated that the less-experienced group was slightly more sensitive than the experienced group, the A' values were not different; collectively, these results indicate that if the groups did actually differ in terms of sensitivity, the difference was not meaningful. The signal detection analyses also revealed that

experienced and less-experienced participants responded conservatively; participants in both groups drew their gun on less than half of the occasions that they were justified in doing so. The β values indicate that, compared to the experienced group, the less-experienced group were substantially more disinclined to draw their gun; the values for c and B'' indicate similar trends.

Table 6
Measures of Sensitivity and Bias by Experience Group

Experience Group	Sensitivity		Bias		
	d'	A'	β	c	B''
Experienced	1.92	.84	9.11	1.15	.87
Less-experienced	2.15	.84	18.17	1.35	.94

Note. d' , β , and c are parametric measures; A' and B'' are non-parametric measures.

Tests of the Experimental Hypotheses

Hypothesis 1: Effect of experience on the ability to predict the actual culminating event. It was hypothesized that experienced participants would predict the actual culminating event more frequently than less-experienced participants. An independent samples t -test demonstrated no significant difference in participants' ability to predict the culminating event, $t(47) = 0.17$, $p = .866$, $d = 0.05$. On average, participants predicted the actual culminating event in slightly more than half of the six test trials ($M_{\text{experienced}} = 3.87$, $SD_{\text{experienced}} = 1.14$; $M_{\text{less-experienced}} = 3.81$, $SD_{\text{less-experienced}} = 1.39$).

It was possible that experienced officers' advantage lay in their ability to predict the actual culminating event as their first assessment option. However, an independent samples t -test also demonstrated no significant difference in participants' ability to predict the actual culminating event first, $t(47) = 0.06$, $p = .949$, $d = 0.02$. The groups'

means were nearly identical to their means for predicting the actual culminating event, indicating that when they predicted the actual culminating event, they typically generated it as their first assessment option.

Hypothesis 2: Effect of experience on the ability to select the criterion

response. It was hypothesized that experienced participants would take the best response option, as identified by the SME, more frequently than less-experienced participants. An independent samples *t*-test revealed that experienced participants ($M = 2.74$, $SD = 1.36$) took the best response option significantly more frequently than less-experienced participants ($M = 1.65$, $SD = 1.16$), $t(47) = 3.02$, $p = .004$, $d = 0.88$. Recall that there were six test trials; amongst those six trials, experienced participants made, on average, one more correct—and potentially life-saving—decision than did the less-experienced participants.

It was also hypothesized that experienced participants would take the best response option—and generate that as their first response option—more frequently than less-experienced participants. An independent samples *t*-test also revealed that experienced participants ($M = 2.52$, $SD = 1.20$) *took the best response option first* significantly more frequently than less-experienced participants ($M = 1.58$, $SD = 1.10$), $t(47) = 2.87$, $p = .006$, $d = 0.84$.

Hypothesis 3: Few assessment/response options generated. It was hypothesized that participants—regardless of experience level—would generate relatively few assessment and response options. On average, participants generated 1.77 ($SD = 0.47$) assessment options and 1.53 ($SD = 0.42$) response options. These numbers

are similar to those reported in other studies that have employed option-generation paradigms (e.g., Johnson & Raab, 2003; Klein & Peio, 1989; Ward et al., 2012).

Hypothesis 4: Relationship between the total number of assessment and response options. It was hypothesized that the total numbers of assessment and response options would be negatively correlated. Contrary to this hypothesis, but consistent with Ward et al. (2011), the total number of assessment options was positively, and significantly, correlated with the total number of response options, $r_s = .50, p < .001$. This appeared to be driven by an underlying significant, positive correlation within the experienced group ($r_s = .48, p = .019$) that was not present in the less-experienced group ($r_s = .24, p = .233$).

Hypothesis 5: Differences in the numbers of task-relevant versus task-irrelevant options generated. It was hypothesized that experienced participants would generate more task-relevant, and fewer task-irrelevant options than less-experienced participants. The mean numbers of task-relevant and task-irrelevant options generated across the six test trials are presented in Figure 1.

Using Wilks's statistic, the 2 (Experience level: experienced, less experienced) by 2 (Type of information: task-relevant option, task-irrelevant option) MANOVA revealed a significant main effect of type of information, $\Lambda = 0.40, F(2, 46) = 34.55, p < .001, \eta_p^2 = .60$. The main effect of skill approached statistical significance, $\Lambda = 0.88, F(2, 46) = 3.02, p = .058, \eta_p^2 = .12$. The Experienced Level \times Information Type interaction was not significant, $\Lambda = 0.91, F(2, 46) = 1.61, p = .211, \eta_p^2 = .07$.

The multivariate analysis was followed up using separate univariate ANOVAs for assessment and response. During assessment, participants—regardless of experience

level—generated significantly more task-relevant ($M = 1.16$, $SD = 0.27$) than task-irrelevant options ($M = 0.61$, $SD = 0.40$), $F(1, 47) = 61.57$, $p < .001$, $\eta_p^2 = .57$. The number of assessment options generated was marginally affected by experience level, $F(1, 47) = 3.57$, $p = .065$, $\eta_p^2 = .07$. The Experience Level \times Information Type interaction was not statistically significant, $F(1, 47) = 0.04$, $p = .843$, $\eta_p^2 = .00$.

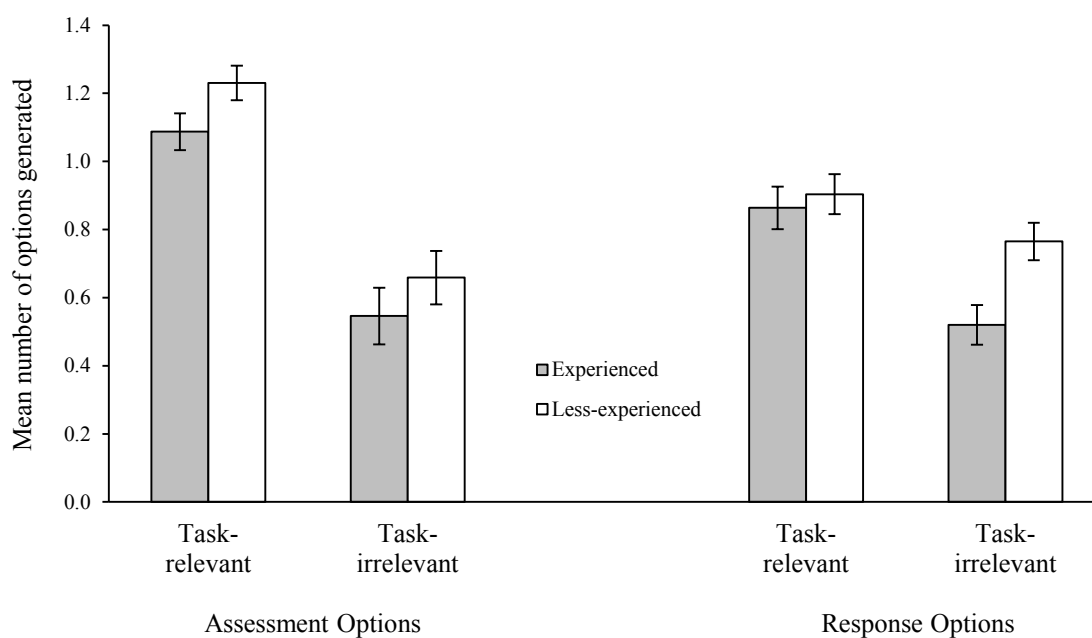


Figure 1. The mean number of task-relevant and task-irrelevant assessment/response options generated, by experience group, for the six test trials. Error bars are $\pm 1 SE$.

For response options, there were significant main effects of experience level and type of information. Experienced participants ($M = 1.38$, $SD = 0.45$) generated significantly fewer response options than less-experienced participants ($M = 1.66$, $SD = 0.36$), $F(1, 47) = 5.67$, $p = .021$, $\eta_p^2 = .11$. Additionally, participants—regardless of experience level—generated significantly more task-relevant ($M = 0.88$, $SD = 0.30$) than task-irrelevant ($M = 0.65$, $SD = 0.30$) options, $F(1, 47) = 16.48$, $p < .001$, $\eta_p^2 = .26$. The

Experience Level \times Information Type interaction was not conventionally significant, $F(1, 47) = 3.10, p = .085, \eta_p^2 = .06$.

This pattern of results is partly consistent with the hypothesis. As expected, participants generated more task-relevant than task-irrelevant options. However, the expected interaction between experience level and type of information was not observed during the assessment phase of decision making, and only approached significance during the response phase.

Hypothesis 6: Threat prioritization of task-relevant assessment options. It was hypothesized that experienced participants would be better than less-experienced participants at prioritizing their task-relevant assessment options. An independent samples t -test demonstrated no significant difference in participants' ability to prioritize their task-relevant assessment options, $t(47) = -0.55, p = .587, d = -0.15$. Recall that the maximum threat prioritization score was 6 points; experienced participants recorded a mean score of 1.90 ($SD = 0.33$), and less-experienced participants recorded a mean score of 1.97 ($SD = 0.54$).

Hypothesis 7: Relationship between the number of task-irrelevant/relevant assessment options and the ability to predict the culminating event. Based on LTWM theory, it was hypothesized that the numbers of task-irrelevant and task-relevant assessment options would be negatively and positively correlated, respectively, with the ability to predict the culminating event. The multiple linear regression indicated that the numbers of task-irrelevant and task-relevant options explained 51.0% of the variance in the number of correct predictions made, $F(2, 46) = 23.93, p < .001$. As expected, the

number of task-irrelevant assessment options was negatively, and significantly, correlated with the number of correct predictions (see Table 7).

Table 7
Summary of Multiple Regression Analysis for Variables Predicting the Number of Correct Predictions Made during Assessment (n = 49)

Variable	<i>B</i>	<i>SE B</i>	β	<i>p</i>
Step 1				
Constant	5.54	0.28		
Task-irrelevant assessment options	-0.18	0.06	-0.34	.002
Task-relevant assessment options	-0.33	0.07	-0.54	<.001
Step 2				
Constant	5.05	0.44		
Task-irrelevant assessment options	-0.19	0.06	-0.36	.002
Task-relevant assessment options	-0.35	0.07	-0.57	<.001
Experience	0.37	0.26	0.15	.165

Note. $R^2 = .51$ for Step 1, $\Delta R^2 = .02$ ($p = .165$).

However, contrary to expectations, the number of task-relevant assessment options was also negatively, and significantly, correlated with the number of correct predictions. Note that the strength of this negative association was larger in magnitude than the negative association between the number of task-irrelevant options and anticipatory performance. The inclusion of experience level as an additional predictor did not significantly improve the model; experience level was not a significant predictor of the number of correct predictions.

Since the relationships between anticipatory performance and the number of task-irrelevant and task-relevant options generated, respectively, were expected to be in opposing directions (i.e., negative and positive, respectively), it was also hypothesized that the total (i.e., task-irrelevant plus task-relevant) number of assessment options would not be significantly correlated with the ability to predict the culminating event. Contrary

to this hypothesis, a regression indicated that the total number of assessment options was negatively, and significantly, correlated with the number of correct predictions made, $R^2 = .49$, $F(1, 47) = 44.37$, $p < .001$, $B = -0.25$, $SE B = 0.04$, $\beta = -0.70$. The inclusion of experience level as an additional predictor did not significantly improve the model, $\Delta R^2 = .02$, $p = .212$, $B = 0.34$, $SE B = 0.27$, $\beta = 0.14$.

Hypothesis 8: Relationship between the number of task-irrelevant/relevant response options and the ability to take the best response option. Based on LTWM theory, it was hypothesized that the numbers of task-irrelevant and task-relevant response options would be negatively and positively correlated, respectively, with the ability to take the best response. The regression indicated that the numbers of task-irrelevant and task-relevant options explained 25.2% of the variance in the number of best response options chosen, $F(2, 46) = 7.75$, $p < .001$. As expected, the number of task-irrelevant response options was negatively correlated with the number of correct responses; however, this relationship was not statistically significant (see Table 8).

Table 8
Summary of Multiple Regression Analysis for Variables Predicting the Number of Times Participants Took the Best Response Option (n = 49)

Variable	<i>B</i>	<i>SE B</i>	β	<i>p</i>
Step 1				
Constant	3.66	0.46		
Task-irrelevant response options	-0.15	0.10	-0.19	.145
Task-relevant response options	-0.30	0.09	-0.42	.002
Step 2				
Constant	4.25	0.58		
Task-irrelevant response options	-0.08	0.10	-0.11	.426
Task-relevant response options	-0.25	0.09	-0.36	.009
Experience	-0.64	0.39	-0.24	.105

Note. $R^2 = .25$ for Step 1, $\Delta R^2 = .04$ ($p = .105$).

However, contrary to expectations, the number of task-relevant response options was negatively, and significantly, correlated with the number of correct responses. The inclusion of experience level as an additional predictor did not significantly improve the model; experience level was not a significant predictor of the number of correct responses.

Based on TTF, it was hypothesized that the total (i.e., task-relevant plus task-irrelevant) number of response options would be negatively correlated with the ability to take the best response option. In support of this hypothesis, a regression indicated that the total number of response options was negatively, and significantly, correlated with the number of times participants took the best response option, $R^2 = .24$, $F(1, 47) = 14.46$, $p < .001$, $B = -0.23$, $SE B = 0.06$, $\beta = -0.49$. The inclusion of experience level as an additional predictor did not significantly improve the model, $\Delta R^2 = .04$, $p = .126$, $B = -0.60$, $SE B = 0.39$, $\beta = -0.22$.

Supplementary analysis 1: Take the first option. On average, participants took (i.e., rated as most likely) their first-generated assessment and response options on 83.0% and 87.4% of test trials, respectively. For assessment, an independent samples *t*-test revealed no significant difference between the frequency with which experienced ($M = 5.09$, $SD = 1.08$) and less-experienced ($M = 4.88$, $SD = 1.14$) participants took their first option, $t(47) = 0.63$, $p = .529$, $d = 0.19$. Similarly for response, there was no significant difference between the frequency with which experienced ($M = 5.35$, $SD = 0.83$) and less-experienced ($M = 5.15$, $SD = 0.73$) participants took their first option, $t(47) = 0.87$, $p = .389$, $d = 0.25$.

Supplementary analysis 2: Entropy. Recall that entropy scores reflect information uncertainty in relation to participants' likelihood ratings: higher entropy scores indicate that participants are more uncertain of which option is most likely. For assessment options, an independent samples *t*-test revealed no significant difference between experienced ($M = 0.48, SD = 0.32$) and less-experienced ($M = 0.64, SD = 0.30$) participants' level of information uncertainty, $t(47) = -1.84, p = .073, d = -0.54$. For response options, experienced participants' ($M = 0.26, SD = 0.26$) had significantly less information uncertainty than the less-experienced participants ($M = 0.49, SD = 0.27$), $t(47) = -3.07, p = .004, d = -0.90$.

Analysis of the CTA Data

Recall that the purpose of conducting the CTA procedure was to collect information that could help provide potential explanations about the differences observed in the option-generation and rating data that could serve as testable hypotheses for the future. As revealed by the tests of the experimental hypotheses, experienced participants' advantage lay in their ability to take the best response option. Therefore, analysis of the CTA data focused on explaining differences in participants' ability to take the best response option.

Also recall that each participant completed the CTA procedure for two of the six test trials, and that those trials were determined, and assigned to each participant, a priori; on average, 16 participants completed the CTA procedure for each test trial. Therefore, the first step was to identify trials that contained sufficient data to permit the most meaningful contrast: experienced and successful participants (i.e., those who took the best response option) compared to less-experienced and unsuccessful participants (i.e.,

those who did not take the best response option). To facilitate this first step, the number of participants in each contrast group was tabulated for each of the six test trials (see Table 9).

Table 9
Number of Participants in Each Contrast Group for the Six Test Trials

Scenario	Contrast group	
	Experienced and successful participants	Inexperienced and unsuccessful participants
Blow up	0	4
Domestic punch	3	4
Gun grab	6	1
Hey Mr Wilson	2	5
Subway	3	1
Suicide waitress	5	3

For each scenario, the number of participants per contrast group was relatively small, and in some cases, unbalanced. The scenarios most suitable for analysis were those that maximized the possibility of identifying a consistent theme within each contrast group—in other words, those with a higher, and balanced, number of participants per contrast group. On this basis, two scenarios—Domestic Punch and Suicide Waitress—were identified as most suitable for analysis; for these scenarios, each contrast group contained at least three participants.

After examining the Suicide Waitress data, it was evident that there was only a minor difference between the contrast groups. Participants in both groups generally indicated that they would have aimed their gun at the waitress, who was about to run at, and attack the officer using a knife; successful participants also indicated that they would shoot the waitress next, while unsuccessful participants did not—at that stage—explicitly

say that they would pull the trigger in the next few seconds. However, the stimulated recall data revealed that the unsuccessful participants were prepared to shoot, and would have shot immediately when the waitress took one more step toward them. In the experimenter's opinion, there were no meaningful differences between participants in these categories; they had all decided to shoot the waitress if she attacked, and there was no evidence that unsuccessful participants were reluctant to defend themselves by shooting the waitress. In light of this, the Suicide Waitress scenario was not subjected to further analysis; a detailed analysis of the CTA data focused on the Domestic Punch scenario. To provide context for the qualitative and quantitative analyses of the retrospective and stimulated recall data, the Domestic Punch scenario is described in Figure 2.

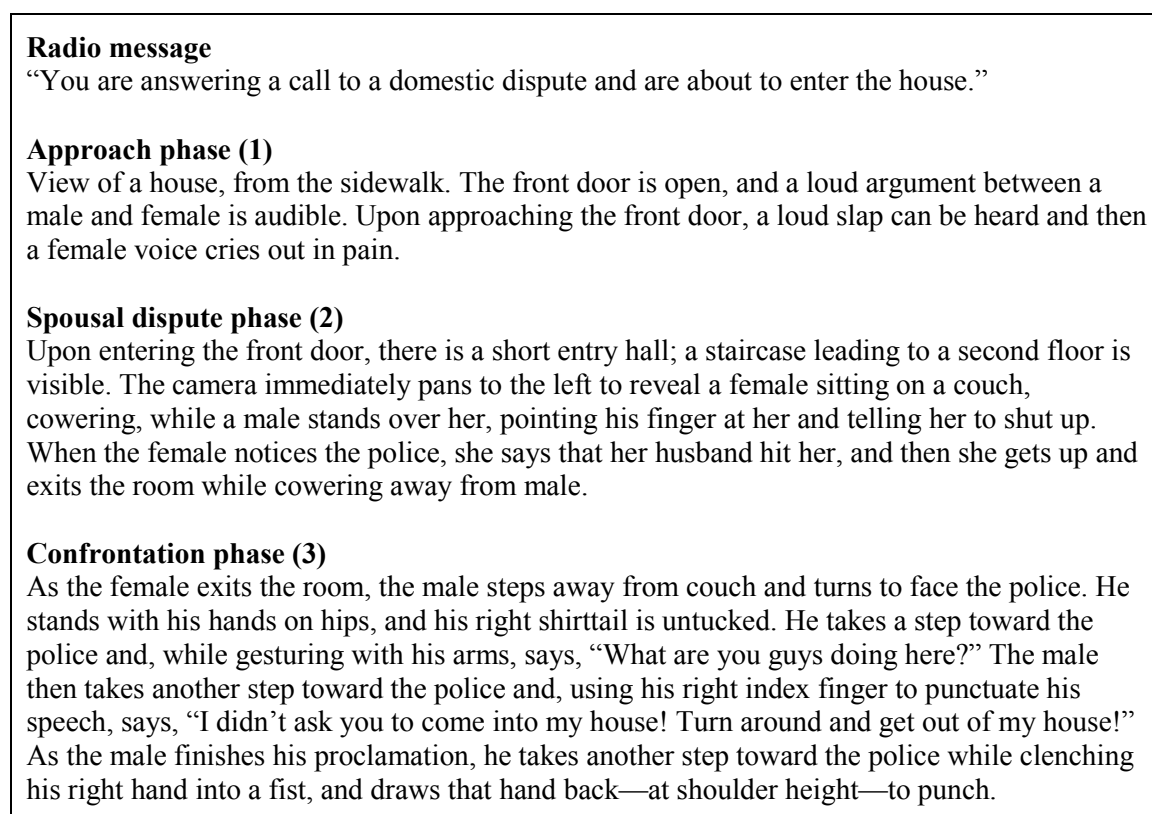


Figure 2. Description of the phases of the Domestic Punch scenario.

Analysis of the retrospective verbal report data (Domestic Punch scenario).

Length of reports. On average, experienced participants' ($M = 83.78$, $SD = 37.22$) retrospective verbal reports contained more words than the reports of less-experienced participants ($M = 71.29$, $SD = 34.96$). This difference was not significant, $t(14) = 0.68$, $p = .505$; however, it did represent a small-sized effect, $d = 0.37$.

Number of statements. On average, experienced participants' ($M = 13.33$, $SD = 5.17$) reports contained more statements than the reports of less-experienced participants ($M = 10.57$, $SD = 6.24$). This difference was not significant, $t(14) = 0.97$, $p = .349$; however, it did represent a medium-sized effect, $d = 0.52$.

Types of statements. The numbers of monitored event, evaluation, inference, prediction, and action statements are described in Table 10.

Table 10
Central Tendency Measures of Statement Types Elicited via Retrospective Verbal Report (Domestic Punch Scenario)

Statement type	Experienced participants ($n = 9$)			Less-experienced participants ($n = 7$)		
	M (SD)	Mdn	Range	M (SD)	Mdn	Range
Monitored event ^a	6.78 (4.44)	7	0–14	5.71 (4.39)	3	2–13
Evaluation ^a	2.22 (1.64)	2	0–5	1.43 (0.98)	1	0–3
Inference ^a	1.78 (0.97)	1	1–3	2.14 (1.68)	2	0–5
Prediction ^b	0.78 (0.97)	1	0–3	1.00 (0.58)	1	0–2
Action ^b	1.00 (1.23)	1	0–3	0.43 (0.79)	0	0–2

^aNormally distributed variable. ^bNon-normally distributed variable.

Independent samples t -tests revealed no significant effect of experience level on the number of cue, $t(14) = 0.48$, $p = .640$, $d = 0.26$, evaluation, $t(14) = 1.13$, $p = .278$, $d = 0.61$, or inference, $t(14) = -0.55$, $p = .592$, $d = -0.29$, statements. Mann-Whitney tests

revealed no significant effect of experience level on the number of predictive inference, $U = 23.00$, $z = -1.01$, $p = .313$, $r = -.25$, or action, $U = 22.50$, $z = -1.06$, $p = .289$, $r = -.27$, statements.

Analysis of the stimulated recall data (Domestic Punch scenario). On average, experienced participants ($M = 4.89$, $SD = 0.78$) stopped the video during the stimulated recall procedure more frequently than did the less-experienced participants ($M = 4.29$, $SD = 1.38$). This difference was not significant, $t(14) = 1.11$, $p = .286$; however, it did represent a medium-sized effect, $d = 0.60$.

Quantitative analysis. Recall that the stimulated recall data were coded with respect to the probe questions, which were designed to elicit information about cues heeded, evaluations/inferences made, prediction/response options generated, and prior knowledge utilized. The median numbers of statements, by statement type and scenario phase, are described in Table 11.

Table 11
Median number of Statements, by Scenario Phase (with duration, in seconds), and Experience Level (Exp. = Experienced, Less-exp. = Less-experienced) for the Domestic Punch Scenario

Statement type	Approach Phase (6.59 s)		Spousal dispute Phase (5.19 s)		Confrontation Phase (7.11 s)	
	Exp.	Less-exp.	Exp.	Less-exp.	Exp.	Less-exp.
Cue heeded	2	2	2	1	3	2
Evaluation	0	0	0	1	1	1
Inference	3	2	1	1	0	0
Prediction option	0	0	0	0	1	1
Response option	1	1	1	1	2	2
Prior knowledge	0	0	0	0	1	0

Overall, participants in each experience group reported similar numbers of each statement type. Note, however, that in the final, confrontation phase, experienced participants heeded more cues and reported more frequent use of prior knowledge than did less-experienced participants.

Qualitative analysis. Recall that the goal of the qualitative analysis was to reveal information that could explain experienced participants' superior ability to take the best response option. The option generation and rating data for the Domestic Punch scenario were used to categorize participants according to their ability to predict the culminating event (i.e., male punches officer) and take the best response option (i.e., strike the male pre-emptively using a punch, kick, or shove) (see Table 12).

Table 12
Categorization of Participants who Completed the Stimulated Recall Task for the Domestic Punch Scenario

Predicted the culminating event	Took the best response option	
	Yes	No
Yes	Experienced ($n = 3$) Less-experienced ($n = 3$)	Experienced ($n = 5$) Less-experienced ($n = 4$)
No	-	Experienced ($n = 1$) Less-experienced ($n = 0$)

In each phase of the scenario, the analysis attempted to identify themes that differentiated between:

1. Experienced and less-experienced participants who predicted the culminating event and took the best response option.

2. Experienced and less-experienced participants who predicted the culminating event but did not take the best response option.
3. Experienced participants who predicted the culminating event/took the best response option and less-experienced participants who predicted the culminating event/did not take the best response option.

The qualitative analysis, therefore, focused on the 15 participants who predicted the culminating event and either did/did not take the best response option. Summaries of those participants' statements during each phase of the Domestic Punch scenario are presented in Tables 13–15. The sole remaining participant—who did not predict the culminating event—was not included in the qualitative analysis, as their data did not offer any additional explanatory power, or serve as a representative point for comparison.

Table 13
Stimulated Recall Statements for Phase 1 (Approach) of the Domestic Punch Scenario by Statement Type, Success in Responding, and Experience Group

Statement type	Predicted and Responded Correctly						Predicted Correctly but Responded Incorrectly								
	Experienced (n = 3)			Less-experienced (n = 3)			Experienced (n = 5)			Less-experienced (n = 4)					
	E1	E2	E3	LE1	LE2	LE3	E4	E5	E6	E7	E8	LE4	LE5	LE6	LE7
Cue headed															
Front door was open	✓	✓	✓	✓	✓	✓			✓				✓		✓
Verbal argument	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓	✓
Slap	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓		✓	✓	✓
Female cried out	✓	✓	✓	✓	✓	✓			✓	✓	✓		✓	✓	✓
Evaluation															
Unusual that front door is open					✓										✓
Entryway/door area is unsafe											✓				
Inference															
Possible presence of weapons			✓		✓			✓							✓
A fight is occurring		✓			✓			✓							✓
At least two people to deal with		✓			✓			✓							✓
Someone was assaulted		✓			✓			✓							✓
Female is in immediate danger		✓			✓			✓							✓
Male is aggressor		✓			✓			✓							✓
No other officer is on-scene															
Officer had cause to enter house									✓						
An arrest needs to be made									✓						✓
Prior history of abuse here															✓
Prediction option															
Argument will continue														✓	
Response option															
Wait outside, yell "Police!"														✓	
Place hand on gun										✓					
Draw gun														✓	
Do not rush into house															
Rush into house		✓											✓		✓
Prior knowledge															
Previous domestic violence calls															✓

Note. Individual participants are labeled by experience group and an identifying number (e.g., E1 = Experienced participant #1, LE7 = Less-experienced participant #7). The cues headed are listed in the order in which they occur in the scenario.

Table 14
Stimulated Recall Statements for Phase 2 (Spousal Dispute) of the Domestic Punch Scenario by Statement Type, Success in Responding, and Experience Group

Statement type	Predicted and Responded Correctly						Predicted Correctly but Responded Incorrectly								
	Experienced (n = 3)			Less-experienced (n = 3)			Experienced (n = 5)			Less-experienced (n = 4)					
	E1	E2	E3	LE1	LE2	LE3	E4	E5	E6	E7	E8	LE4	LE5	LE6	LE7
Cue heeded															
Location is not a kitchen	✓													✓	✓
Female sitting on couch	✓													✓	✓
Male standing over female	✓		✓		✓		✓	✓						✓	✓
Male pointing finger at female		✓			✓									✓	✓
Male shouting at female		✓			✓		✓							✓	✓
Female: "My husband hit me."									✓						
Female holds side of her face		✓		✓		✓								✓	✓
Female gets up, exits room		✓		✓		✓								✓	✓
Evaluation															
Female is upset															
Male is upset with female		✓												✓	
Inference															
Female can leave if she wants	✓														
Female not in immediate danger	✓														✓
Male (aggressor) slapped female			✓		✓		✓	✓						✓	✓
Force needs to be used on male															
No knives in easy reach	✓														
Female is protecting herself															✓
Female is no threat to officer														✓	
Female is out of danger				✓											✓
Prediction option															
Male hits female again															
Female returns to assault officer		✓													
Response option															
Announce police presence														✓	
Scan area for potential weapons			✓												
Separate male and female	✓		✓											✓	
Command male to get on ground														✓	
Talk to male, calm him down															
Restrain male														✓	✓
Prior knowledge															
Training	✓														

Note. Individual participants are labeled by experience group and an identifying number (e.g., E1 = Experienced participant #1, LE7 = Less-experienced participant #7). The cues heeded are listed in the order in which they occur in the scenario.

Table 15
Stimulated Recall Statements for Phase 3 (Confrontation) of the Domestic Punch Scenario by Statement Type, Success in Responding, and Experience Group

Statement type	Predicted and Responded Correctly						Predicted Correctly but Responded Incorrectly								
	Experienced (n = 3)			Less-experienced (n = 3)			Experienced (n = 5)			Less-experienced (n = 4)					
	E1	E2	E3	LE1	LE2	LE3	E4	E5	E6	E7	E8	LE4	LE5	LE6	LE7
Cue heeded															
Male turns to face officer				✓					✓			✓			
Male's right shirt tail is untucked				✓			✓		✓			✓			✓
Male steps toward officer	✓	✓	✓	✓			✓	✓			✓	✓			
Male: "What are you doing here?"				✓					✓						
Male gestures with left hand			✓						✓						
Male puts both hands on hips							✓								
Male points/jabs finger at police				✓											
Male: "Get out of my house!"	✓	✓					✓				✓	✓			✓
Male clenches right hand into fist				✓											
Male raises right hand					✓										
Evaluation															
Male's behavior is aggressive	✓						✓		✓			✓			✓
Male is getting too close to officer							✓								
Male's stance is aggressive								✓							✓
Male's hand gestures are aggressive							✓								✓
Male is non-compliant				✓	✓										✓
Inference															
Male is surrendering	✓														
Shirt tail could be hiding a weapon				✓							✓				
There had been a physical tussle											✓				✓
Male is sizing up the officer				✓											
Officer had cause to enter house											✓				✓
Male is a threat to the officer	✓														
Not enough time to calm male down															
Officer must act pre-emptively	✓	✓	✓	✓	✓	✓									✓
Prediction option															
Male will explain situation to officer	✓														
Male could throw vase at officer				✓											
Male will not be compliant															
Male will punch the officer		✓			✓		✓				✓			✓	
Male is likely to assault officer															✓
Female could return, attack officer				✓											

(Table 15 continues)

(Table 15 continued)

Statement type	Predicted and Responded Correctly						Predicted Correctly but Responded Incorrectly								
	Experienced (n = 3)			Less-experienced (n = 3)			Experienced (n = 5)			Less-experienced (n = 4)					
	E1	E2	E3	LE1	LE2	LE3	E4	E5	E6	E7	E8	LE4	LE5	LE6	LE7
Response option															
Tell male to stop											✓				
Calm male down	✓					✓				✓					✓
Cuff male; seat him on couch															
Scan male for weapons			✓												
Arrest male	✓							✓							
Place hand on gun (in holster)															✓
Prior knowledge															
Use-of-force continuum			✓								✓				✓
Previous incident that was similar						✓				✓					
Spouse is usually main target									✓						
Abusers do not want to go to jail															✓

Note. Individual participants are labeled by experience group and an identifying number (e.g., E1 = Experienced participant #1, LE7 = Less-experienced participant #7). The cues headed are listed in the order in which they occur in the scenario.

To support the goal of explaining experience-level differences in participants' ability to take the best response option in the seconds following the occlusion point, the analysis worked backwards—from the Confrontation phase (3) to the Approach phase (1)—to identify experience-level differences in participants' answers to the probe questions. However, for ease of interpretation, the results are presented in forward chronological order (i.e., starting with phase 1).

Phase 1: Approach. There were no clear differences in participants' answers to the probe questions that facilitated the experience-level comparisons of interest. In general, participants heeded the salient audible cues (e.g., verbal argument, slap, female crying out) and made at least one inference related to an assault-in-progress.

Phase 2: Spousal Dispute. There were no clear differences in participants' answers to the probe questions that facilitated the experience-level comparisons of interest. In general, participants heeded the salient visual cues (e.g., male standing over female, female leaving the room) and typically inferred that the male was aggressor in the situation.

Phase 3: Confrontation. Both experienced and less-experienced participants' ability to predict the culminating event was reflected in their similar responses to the probe questions: they typically heeded the male clenching his fist/raising his hand, and then predicted that the male would assault the officer. Similarities were also observed in the process that led both experienced and less-experienced participants to take the best response option; they all inferred the need for pre-emptive action. Verbatim quotes highlighting this response theme are presented in Table 16.

Table 16
Inferences (underlined) and Response Options that Highlight the Similarity between Experienced and Less-Experienced Participants who Took the Best Response

Experienced participants	Less-experienced participants
E1: <u>“I need to get to him before he gets a chance to do anything further...get him down and handcuffed...he should be detained.”</u>	LE1: “He wasn’t going down without a fight... <u>no other options</u> ...he was going to put my life in danger.”
E2: <u>“I have to take some kind of action against him...either hit him first, or [use a] double-hand shove.”</u>	LE2: <u>“I can’t have that type of aggression in a scene I’m trying to control...I would have probably just clocked him.”</u>
E3: <u>“[I’d] use whatever means necessary...kick him in the crotch... [I] need to get that compliance.”</u>	LE3: “That was a fight, right here... <u>[I’m] just taking him down</u> and handcuffing him.

Note. Participants’ ID names (e.g., E4, LE7) match those reported in Tables 13-15.

Some minor differences were observed in response mode when contrasting experienced and less-experienced participants who predicted the outcome correctly but did not take the best response. Experienced participants typically acknowledged that they may need to engage with the male, physically (i.e., without employing pepper spray, baton, or gun), in order to arrest/detain him. The less-experienced participants, on the other hand, indicated that they would respond using some form of weapon. Verbatim quotes highlighting these trends are presented in Table 17.

Tables 16 and 17 permit a final comparison: A contrast between experienced participants who predicted the outcome correctly and took the best response option (see Table 16, left column), with less-experienced participants who also predicted the outcome correctly but did not take the best response option (see Table 17, right column). As indicated, experienced officers inferred the need to take pre-emptive action against the male, using a strike, or other method (e.g., shove) that did not involve the use of a

weapon. Less-experienced participants, on the other hand, chose to present a weapon (e.g., gun, pepper spray), but did not indicate that they would have used (e.g., shot, sprayed) that weapon. As two of the less-experienced participants reported that they would move backwards (i.e., try to disengage), it is possible that the less-experienced participants were reluctant to engage in a physical scuffle/fight with the male, and thought that presenting a weapon may have deterred him from actually attacking.

Table 17

Experience-based Differences in Response Options Reported by Participants who did not Take the Best Response

Experienced participants	Less-experienced participants
E4: “[I’d] turn him around and cuff him...arrest him.”	LE4: “[I’d] step backwards, draw my gun.”
E5: “He needed to be arrested.”	LE5: “[I’d] have my gun out already.”
E7: “We’re going to be fighting... [I’m] getting ready to go hands on.”	LE6: “I’d attempt to pull out my pepper spray”
E8: “We’re going to fight.”	LE7: “[I’d have my] hand on [my] gun...[while] backing away.”

Note. Participants’ ID names (e.g., E4, LE7) match those reported in Tables 13-15.

Summary

The goal of Study 2 was to further our understanding of the relationship between the situation assessment and response phases of decision making, and to investigate the extent to which this relationship is affected by skill level. An additional goal was to gather information that can be used, in the future, to design training based on the anticipation skills, decision strategies, and processes of experienced police officers.

After viewing temporally-occluded versions of the discriminating scenarios identified in Study 1, participants completed option-generation tasks designed to elicit the

possible situational outcomes and personal courses of action that they were considering at the point of occlusion. These assessment and response “options”, along with their associated ratings (e.g., likelihood, threat) were used to test hypotheses regarding the effect of experience on outcome (i.e., anticipation and decision making) performance, and hypotheses—drawn from two complementary perspectives on option generation—regarding the option-generation processes responsible for superior performance.

With respect to outcome performance, it was expected that experienced participants would anticipate the actual outcome of the scenarios, and select the criterion response, more frequently than the less-experienced participants. The results indicated that although experienced and less-experienced participants did not differ in their ability to anticipate the outcome, experienced participants took the best response option on significantly more of the test trials.

With respect to option-generation processes, it was expected, based on LTWM theory, that experienced participants would generate more task-relevant, and fewer task-irrelevant, options than less-experienced participants. Although the results did not support the expected Experience Level \times Information Type interaction, participants generated more task-relevant than task-irrelevant options, which is generally consistent with LTWM theory. Additionally, compared with less-experienced participants, the experience participants generated significantly fewer options in total during the response phase of decision making; this is consistent with RPD and TTF.

With respect to the relationship between the number of options and outcome performance, LTWM theory posited that there would be a negative correlation between

the number of task-irrelevant options and performance; this relationship was, in fact, observed during both the assessment and response phases of decision making. It was also expected, based on LTWM theory, that the number of task-relevant options would be positively correlated with performance. However, the results revealed that during assessment, the number of task-relevant options was negatively—and not positively—correlated with anticipation performance. During the response phase of decision making, the number of task-relevant options was not significantly correlated with the ability to take the best response option.

Apart from testing the experimental hypotheses, additional analyses were conducted to examine sensitivity and bias in responding, and whether a measure of information uncertainty could differentiate between participants, based on their level of law enforcement experience. Overall, participants were sensitive to the different types of scenarios (i.e., escalating, de-escalating) and responded conservatively: they often did not draw/shoot their gun when they were justified in doing so. Although experienced and less-experienced participants did not differ in their level of information uncertainty during assessment, experienced participants displayed significantly greater certainty in their response options.

To supplement the option-generation and rating data, two CTA techniques were used to provide possible explanations for experienced-based differences in performance, and to gather information that could inform the design of decision-making training in the future. The analysis of the CTA data did, in fact, reveal themes amongst participants' responses that were useful for gaining a deeper insight into cognitive processes and

representations that might have contributed to superior performance. In the Domestic Punch scenario, participants who took the best response option inferred the need to take pre-emptive action (e.g., punch) the suspect. Although experienced participants who did not take the best response option did not infer the need for pre-emptive action, they still recognized that they would need to respond using physical force (i.e., fighting). Less-experienced participants who did not take the best response option indicated that they would draw, but not necessarily use, a weapon. By failing to recognize that the male was not intimidated by the presence of police officers, the less-experienced participants who did not take the best response option placed themselves at a distinct disadvantage: they had drawn a weapon but were not prepared to use it. This left them ill-equipped to deal with the suspect's impending physical attack.

CHAPTER 7

DISCUSSION

Collectively, the purpose of Studies 1 and 2 was to extend the work of Ward et al. (2011) by using a prediction and option-generation paradigm to investigate situation assessment and response selection in dynamic and complex law enforcement situations. Study 1, which was methodological in nature, employed converging approaches to identify video-based law enforcement scenarios that could discriminate between experienced and less-experienced police officers, in terms of their ability to anticipate the culminating event. Study 1 also served to identify a criterion occlusion point in each discriminating scenario: the point that maximized the experienced-based difference in anticipation performance. Study 2 then employed those discriminating scenarios, and their optimal occlusion points, to test hypotheses related to experience-based performance differences and the option generation processes proposed to be responsible for superior performance. Recall that two main perspectives on option generation were contrasted: LTWM theory, which has previously been tested during the situation assessment phase of decision making, and RPD/TTF, which focuses on the response selection phase of decision making.

In general, LTWM theory posits that superior performers generate a detailed, up-to-date, and accurate representation of the situation, and that this facilitates their ability to anticipate the situational outcome. Superior anticipation performance should, therefore, be associated with the generation of more task-relevant, but fewer task-irrelevant assessment options (Ward et al., 2012). Additionally, superior anticipation performance

should be associated with the ability to prioritize task-relevant assessment options according to the threat that they pose (Ward et al., 2012). Accordingly, LTWM theory suggests that when multiple task-relevant assessment options exist in dynamic, uncertain, and time-constrained environments, often times there should be a positive relationship between the number of those options generated and anticipation performance—depending, of course, on the utility of that mechanism in that environment. In other words, perceiving more of the threatening options, while ignoring irrelevant information, should lead to better anticipation performance. As a corollary, Ward et al. (2011) expected that the ability to generate more task-relevant options during assessment that describe the semantic relations between situational information cues—to paint a detailed picture of the situation—should facilitate use of a TTF-like mechanism during response selection. That is, the ability to anticipate the situational outcome should enable skilled decision makers to generate, as their first response option, a satisfactory, if not a very good, response.

With respect to the response phase of decision making, TTF posits that the number of response options generated is negatively correlated with the quality of the final decision. In other words, the ability to select a good response results from generating fewer, rather than more response options, which according to TTF would be better options, thus facilitating selection of one of the better, earlier options generated. However, while LTWM theory makes the distinction between task-relevant and task-irrelevant options (i.e., during situation assessment), TTF makes no such distinction; the

hypothesized negative relationship is between the total number of response options generated, and the quality of the final decision.

Outcome Performance

Hypothesis 1 was that experienced participants would anticipate the culminating event more frequently than less-experienced participants. Although, on average, both experienced and less-experienced participants anticipated the culminating event in approximately four of the six test trials, they did not differ significantly in their anticipation ability. This was surprising, given that the stimuli and occlusion points employed were those identified as discriminating of experience level in Study 1. Furthermore, this finding failed to support one of the most consistent findings in the expertise literature: that superior anticipation skill is a hallmark of superior performance. One possible explanation for this finding is that despite their wealth of practical experience, the experienced participants in Study 2 had not developed reliably superior perceptual anticipation. Although this is concerning—in terms of officers' safety—it does point to a need for improving police officers' anticipation skills. In the sport domain, anticipation training based on the perceptual skills of expert performers has led to improvements in perceptual anticipation (e.g., Smeeton, Williams, Hodges, & Ward, 2005). Similar evidence-based training could also prove effective in the law enforcement domain. An alternative explanation is that the scenarios used in the research were not representative of the knowledge structures that police officers develop. However, counter to this explanation, 50% of the lethal scenarios used in Study 2 have been shown to discriminate between highly trained SWAT officers and police recruits.

Hypothesis 2 was that experienced participants would take the best response option more frequently than less-experienced participants. Consistent with this hypothesis, experienced participants were better at generating the criterion response option, and then rating that option as the one they were most likely to pursue. It is interesting that experienced participants displayed superior response performance in the absence of superior anticipatory skills. One possible explanation for this is that experienced participants have better mapping between representative situations and optimal responses. For example, they may realize that, when faced with a suspect carrying an improvised explosive device (e.g., Blow Up scenario), the optimal solution is to shoot the suspect immediately, before the suspect can detonate the device. Although less-experienced participants may also consider this course of action, they typically pursue a different, but seemingly good, response (e.g., moving innocent bystanders away from the suspect) that will not actually resolve the problem at hand. The notion that experienced operators recognize familiar situations, and that recognition primes an associated, appropriate response, is one of the central tenets of the RPD model. However, it is worth noting that this is not a feature unique to the RPD model; consistent with LTWM theory, experienced operators' ability to index information at encoding, in conjunction with their superior situation-model building skills, would also encompass appropriate responses. With respect to the issue of mapping responses to situations, RPD and LTWM theory differ in the number of task-relevant response options that are activated: a single, satisfactory option in the "optimized" version of RPD (see Johnson & Raab, 2003; Yates, 2001), versus all task-relevant response options (i.e., where more than

one exists), in the case of LTWM theory. To continue the Blow Up scenario example, a rich situation model may include the possible responses of “Shoot the suspect in the chest” and “Shoot the suspect in the head.” Although both responses are satisfactory, shooting the suspect in the head is the best option (i.e., when the situation allows), because it typically results in instant incapacitation, whereas a shot to the body may still allow the suspect to detonate the device. Having both options available—in the context of a rich situation model—would allow an officer to make the best decision given the specific circumstances (e.g., distance from suspect, the officer’s shooting accuracy, the possibility of hitting an innocent bystander if the shot misses). Clearly, a person’s ability to represent and act on the environmental structure dramatically influences their skill in making use of the most appropriate strategy. The importance of a systems perspective (i.e., consideration of person, process, and environment) when evaluating superior decision making has been discussed elsewhere (e.g., Cokely & Kelley, 2009; Ward et al., 2011), and is likely applicable here.

Option-Generation Measures

Based on previous option-generation studies, Hypothesis 3 was that participants would generate few (i.e. less than 5) assessment and response options. Consistent with this hypothesis, participants generated, on average, slightly less than two assessment/response options per trial.

Hypothesis 4 was that the total numbers of assessment and response options would be negatively correlated. This was based on the notion that possessing a detailed situation model (i.e., quantified by the number of assessment options generated) would

facilitate a TTF-like mechanism during the response phase of decision making (see Ward et al., 2011). However, the data did not support this hypothesis. In fact, contrary to the hypothesis, experienced participants exhibited a positive correlation between the numbers of assessment and response options generated; no significant correlation—either negative or positive—was found for less-experienced participants. Note that this finding—regarding the observed positive correlation—corresponds with the finding of Ward et al. (2011), and that both findings provide evidence against the hypothesized relationship. One possible reason for the experienced participants' positive correlation is that they tend to match each assessment option with a single response. Such behavior would be advantageous in cases where the anticipated situational outcomes differ in terms of the level-of-threat they present. For example, perceived low-level threats could be dealt with by issuing strong verbal commands, whereas perceived high-level threats may necessitate the use of a weapon.

Hypothesis 5 related to differences in the number of task-relevant and task-irrelevant options generated. Recall that from the perspective of LTWM theory, skilled performers construct accurate and up-to-date situation models, which aid in identifying meaningful patterns of cues in the environment, while disregarding irrelevant information. Therefore, it was expected that experienced participants would generate more task-relevant than task-irrelevant assessment options, and that they would generate fewer task-irrelevant assessment options than less-experienced participants. The results of Study 2 provide partial support for this hypothesis. As expected, experienced and less-experienced participants generated significantly more task-relevant than-irrelevant

assessment options; however, compared to the less-experienced participants, experienced participants did not generate significantly fewer task-irrelevant assessment options.

Although this hypothesis, based on LTWM theory, primarily relates to assessment options, recall that Ward et al. (2011) also found the hypothesized relationship when they compared the numbers of task-relevant and task-irrelevant response options. Although the current study did not find a significant interaction between experience level and type of information (i.e., task-relevant versus task-irrelevant), experienced participants generated significantly fewer response options than less-experienced participants, and in general, participants generated more task-relevant than task-irrelevant response options.

To summarize, consistent with LTWM theory, participants generated more task-relevant than task-irrelevant options. There was, however, no evidence found to support the contention that experienced participants would demonstrate a greater reduction in the amount of task-irrelevant information generated, compared to less-experienced participants. Keeping in mind that relatively few task-irrelevant options were generated at all, it is possible that a key difference between skilled and less-skilled performers lies not in the number of task-irrelevant options generated, but in the way those options are contextualized within the set of options generated. For example, a skilled performer might—if they generate a task-irrelevant option—be better at discounting that option (e.g., rating it as an unlikely situational outcome/not a good course of action), whereas less-skilled performers may fail to recognize the same option as task-irrelevant, and could, therefore, fail to discount it. Given the dynamic and uncertain nature of law enforcement situations, generating some task-irrelevant options may be inevitable; the

ability to reduce any detrimental effect that those options have on perceptual anticipation and selecting a course of action could be an important determinant of skilled performance.

Based on LTWM theory, Hypothesis 6 was that experienced participants would be better than less-experienced participants at prioritizing their task-relevant assessment options. The results, however, showed that there was no significant experience-based difference in threat prioritization. This can be explained by considering several factors. First, on average, participants generated slightly more than one task-relevant assessment option per trial. Second, task-irrelevant options were typically rated by the SME as posing less (i.e., and not more) of a threat than the task-relevant options. Third, the culminating event was often the most threatening, task-relevant assessment option. Fourth, there was no significant difference between the groups in terms of their anticipation performance. In light of these factors, it appears that participants in both groups often predicted the culminating event (i.e., the most threatening task-relevant option) and recognized it as a highly threatening outcome. In this case, and in the absence of other task-relevant options, it appears that there was relatively little opportunity for the prioritization scores to vary. Experienced-based differences in participants' ability to prioritize task-relevant options may only become apparent in situations with greater uncertainty in the situational outcome. Under such conditions, participants might generate more task-relevant options to cover the various, plausible outcomes, thus increasing the potential for variability when ranking those options by threat.

Hypothesis 7 was that the numbers of task-irrelevant and task-relevant assessment options would be negatively and positively correlated, respectively, with the ability to predict the culminating event. As expected, the number of task-irrelevant assessment options was negatively associated with the ability to predict the culminating event; participants who generated fewer task-irrelevant options across the six test trials predicted the culminating event in more of those trials. With respect to LTWM theory, this finding supports the contention that superior performers construct situation models that contain less irrelevant information. LTWM theory also posits that superior performers construct more detailed situation models. One way in which the level of detail contained within a situation model can be measured is via the number of task-relevant options generated. However, contrary to expectations, the number of task-relevant assessment options was negatively, rather than positively, correlated with the ability to predict the culminating event. Recall that for this analysis, the number of task-relevant assessment options did not include the culminating event if it was generated by participants. Therefore, the negative correlation observed between the number of task-relevant options and the number of correct anticipations suggests that other than generating the culminating event, there was no benefit gained by generating additional task-relevant options.

One explanation for this finding is that by the time the scenario was occluded, participants who predicting the culminating event envisioned very few, or even only one, way in which the scenario could culminate. If this was the case, a question can be asked: Do participants ever, in the evolution of a single scenario, generate more of the task-

relevant options present in the environment, or do they typically only generate the task-relevant option they perceive as the one most likely to occur? Although Study 2 did not address this question directly, the stimulated recall data indicate that participants, typically, did not generate many predictions prior to the occlusion point and did not generate more than one prediction each time they stopped video playback. Such behavior during situation assessment is more consistent with Klein's (1989) description of situation assessment, in which operators typically generate, and then consider, only one explanation (i.e., assessment) of the situation at any given time. Klein reported that in situations where multiple assessment options were generated, skilled operators generally evaluated those options in a serial manner; less-skilled operators were more inclined to evaluate multiple options concurrently (i.e., by contrasting/deliberating between options).

Hypothesis 8 was related to response options: according to LTWM theory, it was expected that the numbers of task-irrelevant and task-relevant response options would be negatively and positively correlated, respectively, with the ability to take the best response. The results showed that, as expected, the number of task-irrelevant response options was negatively associated with the ability to take the best response option; participants who took the best response more frequently also generated fewer task-relevant options. Therefore, it appears that reducing the amount of irrelevant information (i.e., task-irrelevant response options) facilitated superior decision making, just as reducing the amount of task-irrelevant information during assessment facilitated perceptual anticipation. This suggests that, at least to some degree, similar option-generation mechanisms are at play during assessment and response.

However, contrary to the expected, positive relationship between the number of task-relevant response options and the number of times participants took the best response, no significant relationship—either positive or negative—was observed. This suggests that as long as participants generated the best response option, generating additional, task-relevant response options did not have a detrimental effect on their performance.

Juxtaposed to LTWM theory's predictions regarding the relationship between the number of response options and decision-making performance, TTF predicts an overall negative relationship between the total number of response options and the ability to take the best response option. The results supported this prediction: participants who took the best response more frequently also tended to generate fewer response options. In the absence of any significant relationship between the number of task-relevant response options and decision-making performance, it appears that the overall negative correlation was driven by the significant negative correlation observed for the number of task-irrelevant options. Although this result—at face value—provides support for TTF, TTF posits that the negative relationship is due to a decrease in option quality with serial position. However, an alternative mechanism that could explain the negative relationship is an overall reduction in the amount of task-irrelevant information considered. This is consistent with Haider and Frensch (1996), who suggested that over time, operators learn to distinguish between task-relevant and task-irrelevant information, gradually becoming better at focusing on the relevant information while restricting the extent to which they process irrelevant information.

Supplementary Points

Signal Detection Performance

To determine whether participants were sensitive to differences in the types of scenarios (i.e., whether a “draw gun” response was justified or unjustified), and whether they were biased to respond by drawing/shooting their gun, a signal detection analysis was performed. The analysis found that both experienced and less-experienced participants were highly sensitive to the different types of scenarios; there were exceedingly few instances in which a participant drew/shot their gun when they were not justified in doing so. This is good news for police departments, and for the public-at-large. On the other hand, both experienced and less-experienced participants showed a strong conservative bias in using their gun: when officers were justified in drawing/shooting their gun, they often did not do so. This actually presents an officer-safety problem, because there are some—albeit potentially low-frequency—situations in which an officer’s best chance of survival involves the use of lethal force. Although adopting a conservative response policy reflects a public-, rather than a self-, preservation strategy, ideally officers should be perfectly calibrated in terms of their response in order to maximize their effectiveness: they should draw their gun only when necessary, but they should also draw their gun every time it is necessary to do so. One potential way of reducing this conservative bias toward a more unbiased response policy is by exposing police officers, during training, to high-threat/low-frequency situations which demand a lethal response such that they can be exposed to, and could integrate, the full range of ecologically valid cues (see Brunswik, 1957) into their existing situational representation.

Of course, such training would need to ensure that officers attain an unbiased policy (i.e., that they do not overcompensate and adopt a liberal response bias).

Taking the First-Generated Option

Consistent with TTF, participants took their first-generated response option a majority of the time; participants also adopted a TTF strategy during assessment. Here, it is worth pointing out that both TTF and LTWM theory are both activation-based models, and therefore, both models would predict that better options are generated earlier in the option generation sequence (Johnson & Raab, 2003; Ward et al., 2012). Assuming that all else is equal with respect to the environment, one major difference between the two perspectives in the specific environment tested is that TTF holds the best decision will be made by taking the first-generated option; LTWM theory holds that the best option is likely to be amongst the first few options generated, but not necessarily the first.

Entropy/Information Uncertainty

Regarding information uncertainty, experienced participants displayed less uncertainty in their assessment options than did the less-experienced participants. Although the difference was not statistically significant, it constituted a medium-sized effect. Given that experienced and less-experienced participants generated similar numbers of assessment options, the difference in information uncertainty is attributable to differences in the weighting of assigned likelihood ratings. In other words, when experienced participants generated more than one assessment option, they rated one option as much more likely than the other (e.g., 90% versus 10%). Less-experienced

participants, on the other hand, tended to rate their assessment options more similarly (e.g., 60% versus 40%), thus displaying greater information uncertainty.

Although experienced participants expressed significantly less information uncertainty for their response options than did less-experienced participants, another factor—other than the weighting of assigned likelihood ratings—may have influenced this effect. Recall that the analysis of task-relevant versus task-irrelevant response options found a main effect of experience level: experienced participants generated significantly fewer response options overall than did less-experienced participants. Recall too that information uncertainty is reduced by generating fewer options. Therefore, the reduction in information uncertainty for response options was also due, at least in part, to the reduction in the number of options generated.

Although measures of information uncertainty could potentially be used to predict a participant's experience level, care should be taken with regard to interpreting entropy. Based on the data from Study 2, it would appear that an absolute reduction in information uncertainty (i.e., toward $H = 0$) is associated with better performance. From the perspective of TTF, if the first option generated is typically the best option, then reducing uncertainty toward $H = 0$ makes sense: generate a single option that, by default, must be rated as 100% likely. However, from the perspective of LTWM theory, an absolute reduction in uncertainty is not advisable, since generating a single option means that other, task-relevant options will be omitted from the option set, potentially limiting the extent to which individuals could adapt their behavior on-the-fly. Therefore, from a LTWM-theory perspective, any reduction in information uncertainty should take place

within the context of the number of task-relevant options that could be generated in a given situation (i.e., a relative, rather than an absolute reduction). For example, consider a situation in which there are three task-relevant options, and where one of those options should be considered as “very likely” to occur next, while the other two are quite unlikely to occur next. A completely uncertain participant would rate all three options as equally uncertain (i.e., all 33.3% likely); the information score, H , would equal 1.58, and such behavior would not be indicative of skilled performance. On the other hand, rating one of the three options as very likely and the other two as unlikely (e.g., 90%, 8%, and 2%) results in $H = 0.54$. The second situation, representing greater certainty, is associated with a relatively low (i.e., to complete uncertainty, $H = 1.58$) entropy score, and would, in the situation described, represent skilled performance in terms assigning accurate likelihood ratings.

Limitations and Future Research Directions

The lack of support for a number of the hypotheses may, in part, be attributable to methodological issues. For example, the fact that there was no experience-based difference in anticipation performance may be due to sampling issues in Study 1. Recall that Study 1 was designed to identify occlusion points that maximized the difference between experienced and less-experienced participants, in terms of their ability to anticipate the culminating event. Although every attempt was made to ensure a substantial difference in experience between the groups within Studies 1A–C, the extent of this difference varied across the studies. It should also be noted that each study sampled participants from a different population (e.g., large, mainly urban department

versus small, mainly rural department); this is likely to have had some effect on the frequency and types of incidents that officers encounter and train for. Furthermore, the literature on expert performance highlights that measures such as level-of-experience, rank, and status, are not necessarily indicative of actual performance. Currently, there are no reliable methods for assessing the skill of police officers in complex and dynamic law enforcement situations. One potential avenue for further research is to first define what constitutes skilled performance in these situations, and then develop assessment and training tools that have criterion validity. Another avenue may be to examine the extent to which these domain-specific skills and processes have greater predictive validity than domain-general measures of cognitive ability (e.g., intelligence, working memory capacity, spatial ability, risk literacy) (see Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012; Daneman & Carpenter, 1980; Peters, Laeng, Latham, Jackson, Zaiyouna, & Richardson, 1995; Raven, 2000). All this said, the current data may suggest that rather than being a consequence of superior prediction skills, the important source of skill differences in law enforcement may be related to how individuals interpret and translate situational information into an effective response. More research is needed to examine this further.

Another factor that may have affected the results—in particular, the number of options generated—was the process via which options were elicited. First, it must be noted that in general, law enforcement training is typically centered on officers' ability to respond correctly. Therefore, officers intuitively wanted to report a single response option: what they would do next, rather than what they could do next. Furthermore, some

participants found the assessment-option generation task quite challenging, because in their minds, they would have already taken action prior to the occlusion point that may have changed the outcome if they were in an actual, rather than a simulated, situation. In other words, their actions—carried out in their minds—may have constrained the situation, and impacted the number and type of options they generated.

An alternative way of assessing theories related to option generation could be through the use of an option selection and ranking paradigm (e.g., Hintze, 2008). However, use of such a paradigm involves a tradeoff: The fact that participants select pre-specified options from a list may lead them to select more options than they would have otherwise generated, thus providing invalid results. The validity and utility of such a paradigm must be assessed empirically, and should be compared with data gathered via option-generation tasks.

Recall that one of the goals of this dissertation was to gather information that could be used, in the future, to design training based on the anticipation skills, decision strategies, and processes of experienced officers. The fact that participants were able to anticipate the outcome in only four of the six test trials suggests that both experienced and less-experienced officers could benefit from perceptual anticipation training. Alternatively, training that is geared toward matching their encoding of the potential threats to the selection of better response options may prove equally, if not more, beneficial. Given that the number of task-irrelevant options was negatively correlated with anticipation and decision-making performance, officers may also benefit from

training that improves their ability to distinguish between task-relevant and task-irrelevant options.

In the USA, many states have developed training curricula that must be followed by all police academies. These curricula specify, in detail, the amount of time to be spent on each topic. As a result, there is often little, or no, opportunity to administer additional training beyond that which is mandated. In light of this, any anticipation and/or decision making training for law enforcement should aim to minimize the amount of time and resources needed for delivery. One possible way to overcome obstacles to delivering such training could involve the development of online learning and testing software that recruits can complete individually, at their own leisure. Such a system may also be of benefit to law enforcement departments, which typically have to cover many topics (i.e., legal updates, weapon requalification) during annual, in-service training.

Summary

Studies 1 and 2 represent a principled approach to understanding situation assessment and decision making in law enforcement. In terms of outcome performance, experienced participants were significantly better than less-experienced participants at taking the best response option, but not at anticipating the situation outcome. It is not yet clear whether experienced-based differences in perceptual anticipation performance exist in the law enforcement domain. However, performance data from a previous study (Ward et al., 2011) suggests that such perceptual skill does exist. One challenge, then, is to identify the conditions under which this skill can be detected reliably. Once such skill is identified, future research can assess whether it can be improved through training.

Two main perspectives regarding the option generation processes that lead to superior performance were examined: TTF and LTWM theory. In general, participants generated few assessment and response options, and consistent with TTF, typically rated their first-generated option as the most likely option. One major finding was that, consistent with LTWM theory, participants generated fewer task-irrelevant options than task-relevant options, and the number of task-irrelevant options generated was negatively associated with anticipation and decision-making performance. This suggests that training participants to distinguish task-relevant from task-irrelevant information could result in improved performance.

However, contrary to the prediction made by LTWM theory, the number of task-relevant assessment options was negatively, and not positively, correlated with anticipation performance. This result was more consistent with a TTF-like mechanism, and suggests that there are some situations in which it is advantageous to generate fewer, and not more, of the task-relevant options that are available in the environment. This finding highlights the importance of defining the boundary conditions under which different types of option generation behavior are likely to be adaptive.

In general, this research has increased our understanding of situation assessment and decision making in a dynamic and complex domain, and provides a foundation on which to continue investigating the basis of skilled performance in law enforcement, with a view to designing training that will save lives. The methods used here can be applied to deepen the understanding of these processes in other domains.

CHAPTER 8**REFERENCES**

- Abernethy, B. (1987). Anticipation in sport: A review. *Physical Education Review, 10*, 5–16.
- Abernethy, B. (1990). Anticipation in squash: Differences in advance cue utilization between expert and novice players. *Journal of Sports Sciences, 8*, 17–34. doi: 10.1080/02640419008732128
- Abernethy, B., & Russell, D. G. (1987). Expert–novice differences in an applied selective attention task. *Journal of Sport Psychology, 9*, 326–345.
- Adelman, L., Gualtieri, J., & Stanford, S. (1995). Examining the effect of causal focus on the option generation process: An experiment using protocol analysis. *Organizational Behavior and Human Decision Processes, 61*, 54–66. doi: 10.1006/obhd.1995.1005
- Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley.
- Brezovic, C. P., Klein, G. A., & Thordsen, M. (1990). *Decision making in armored platoon command* (ARI Research Note 90-51). Alexandria, VA: United States Army Research Institute for the Behavioral and Social Sciences.
- Brunswik, E. (1957). Scope and aspects of the cognitive problem. In J. S. Bruner, E. Brunswik, L. Festinger, F. Heider, K. F. Muenzinger, C. E. Osgood & D. Rapaport (Eds.), *Contemporary approaches to cognition* (pp. 5–31). Cambridge, MA: Harvard University Press.
- Calderwood, R., Crandall, B. W., & Baynes, T. H. (1990). *Protocol analysis of*

expert/novice command decision making during simulated fire ground incidents

(ARI Research Note 90-44). Alexandria, VA: United States Army Research Institute for the Behavioral and Social Sciences.

Calderwood, R., Crandall, B. W., & Klein, G. A. (1987). *Expert and novice fire ground command decisions* (KATR-858(D)-87-02F). Yellow Springs, OH: Klein Associates Inc.

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith (Eds.), *Towards a general theory of expertise: Prospects and limits* (pp. 195–217). New York, NY: Cambridge Press.

Chabris, C. F., & Hearst, E. S. (2003). Visualization, pattern recognition, and forward search: Effects of playing speed and sight of the position on grandmaster chess errors. *Cognitive Science*, 27, 637–648. doi: 10.1016/S0364-0213(03)00032-6

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37–46. doi: 10.1177/001316446002000104

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Psychology Press.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, 7, 25–47.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and*

Decision Making, 4, 20–33.

Crandall, B. W., & Getchell-Reiter, K. (1993). Critical decision method: A technique for eliciting concrete assessment indicators from the intuition of NICU nurses.

Advances in Nursing Science, 16, 42–51.

Crandall, B. W., Klein, G. A., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. Cambridge, MA: MIT Press.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466. doi: 10.1016/s0022-5371(80)90312-6

de Groot, A. D. (1965). *Thought and choice in chess* (1st ed.). The Hague: Mouton Publishers.

Doane, S. M., Sohn, Y. W., & Jodlowski, M. T. (2004). Pilot ability to anticipate the consequences of flight actions as a function of expertise. *Human Factors*, 46, 92–103. doi: 10.1518/hfes.46.1.92.30386

Durso, F. T., Bleckley, M. K., & Dattel, A. R. (2006). Does situation awareness add to the validity of cognitive tests? *Human Factors*, 48, 721–733. doi: 10.1518/001872006779166316

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly*, 6, 1–20.

Durso, F. T., Rawson, K., & Giroto, S. (2007). Comprehension and situation awareness. In F. T. Durso, R. Nickerson, S. Dumais, S. Lewandowsky & T. Perfect (Eds.),

Handbook of applied cognition (2nd ed., pp. 163–193). Chichester, UK: John Wiley & Sons.

Durso, F. T., & Sethumadhavan, A. (2008). Situation awareness: Understanding dynamic environments. *Human Factors*, *50*, 442–448. doi: 10.1518/001872008x288448

Endsley, M. R. (1988, May). *Situation awareness global assessment technique (SAGAT)*. Paper presented at the Proceedings of the National Aerospace and Electronics Conference, New York.

Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*, 32–64. doi: 10.1518/001872095779049543

Endsley, M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*, 65–84. doi: 10.1518/001872095779049499

Endsley, M. R. (2000). Situation models: An avenue to the modeling of mental models. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *44*, 61–64. doi: 10.1177/154193120004400117

Endsley, M. R. (2006). Expertise and situation awareness. In K. A. Ericsson, N. Charness, P. J. Feltovich & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 633–651). New York: Cambridge University Press.

Endsley, M. R., Sollenberger, R., Nakata, A., & Stein, E. S. (1999). *Situation awareness in air traffic control: Enhanced displays for advanced operations* (DOT/FAA/CT-TN00/01). Atlantic City, NJ: Federal Aviation Administration William J. Hughes Technical Center.

- Engelmann, P. D., & Gettys, C. F. (1985). Divergent thinking in act generation. *Acta Psychologica, 60*, 39-56. doi: 10.1016/0001-6918(85)90012-5
- Ericsson, K. A., Charness, N., Feltovich, P., & Hoffman, R. R. (Eds.). (2006). *The Cambridge handbook of expertise and expert performance*. Cambridge, UK: Cambridge University Press.
- Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102*, 211–245. doi: 10.1037//0033-295X.102.2.211
- Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual Review of Psychology, 47*, 273–305. doi: 10.1146/annurev.psych.47.1.273
- Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (revised ed.). Cambridge, MA: Bradford books/MIT Press.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191. doi: 10.3758/BF03193146
- Federal Bureau of Investigation. (2010a). *Law enforcement officers feloniously killed: Type of weapon, 2000–2009* [Data file]. Retrieved from <http://www2.fbi.gov/ucr/killed/2009/data/documents/table27.xls>
- Federal Bureau of Investigation. (2010b). *Crime in the United States, 2009: Full-time law enforcement employees by population group, percent male and female* [Data file]. Retrieved from <http://www2.fbi.gov/ucr/cius2009/data/documents/09tbl74.xls>
- Federal Bureau of Investigation. (2012). *Law enforcement officers feloniously killed:*

Region, geographic division, and state, 2002–2011 [Data file]. Retrieved from <http://www.fbi.gov/about-us/cjis/ucr/leoka/2011/LawEnforcementOfficersKilledandAssaultedTables.zip> (see Table 1)

Gettys, C. F., & Fisher, S. D. (1979). Hypothesis plausibility and hypothesis generation.

Organizational Behavior and Human Performance, 24, 93–110. doi:

10.1016/0030-5073(79)90018-7

Gettys, C. F., Pliske, R. M., Manning, C., & Casey, J. T. (1987). An evaluation of human

act generation performance. *Organizational Behavior and Human Decision*

Processes, 39, 23–51. doi: 10.1016/0749-5978(87)90044-6

Gigerenzer, G., & Brighton, H. (2009). Homo heuristics: Why biased minds make better

inferences. *Topics in Cognitive Science*, 1, 107–143. doi: 10.1111/j.1756-

8765.2008.01006.x

Gigerenzer, G., & Todd, P. M., & the ABC Research Group (1999). *Simple heuristics*

that make us smart. New York, NY: Oxford University Press.

Gobet, F., & Simon, H. A. (1996a). Templates in chess memory: A mechanism for

recalling several boards. *Cognitive Psychology*, 31, 1–40. doi: 10.1006/cogp.

1996.0011

Gobet, F., & Simon, H. A. (1996b). The roles of recognition processes and look-ahead

search in time-constrained expert problem solving: Evidence from grand-master-

level chess. *Psychological Science*, 7, 52–55. doi: 10.1111/j.1467-9280.1996.

tb00666.x

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The

recognition heuristic. *Psychological Review*, *109*, 75–90. doi: 10.1037//0033-295X.109.1.75

Goodman, L. A. (1954). Kolmogorov-Smirnov tests for psychological research. *Psychological Bulletin*, *51*, 160–168.

Hagemann, N., Schorer, J., Canal-Bruland, R., Lotz, S., & Strauss, B. (2010). Visual perception in fencing: Do the eye movements of fencers represent their information pickup? *Attention, Perception, & Psychophysics*, *72*, 2204–2214. doi:10.3758/APP.72.8.2204

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology*, *30*, 304–337. doi: 10.1006/cogp.1996.0009

Harris, K. R., Tashman, L., Ward, P., Ericsson, K. A., & Eccles, D. W. (2006). Planning, evaluation, and cognition: Exploring the structure and mechanisms of expert performance in a representative dynamic task. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *28*, 327–332.

Haskins, M. J. (1965). Development of a response-recognition training film in tennis. *Perceptual and Motor Skills*, *21*, 207–211.

Helsen, W., & Pauwels, J. M. (1993). The relationship between expertise and visual information processing in sport. In J. L. Starkes & F. Allard (Eds.), *Cognitive issues in motor expertise*. (pp. 109–134). Amsterdam, Netherlands: North-Holland.

Hintze, N. (2008). *First responder problem solving and decision making in today's asymmetrical environment* (Doctoral dissertation, Naval Postgraduate School).

Retrieved from <http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA479926>

- Hogan, M. P., Pace, D. E., Hapgood, J., & Boone, D. C. (2006). Use of human patient simulation and the Situation Awareness Global Assessment Technique in practical trauma skills assessment. *Journal of Trauma-Injury Infection and Critical Care*, *61*, 1047–1052. doi: 10.1097/01.ta.0000238687.23622.89
- Hogg, D. N., Follesø, K., Strand-Volden, F., & Torralba, B. (1995). Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms. *Ergonomics*, *38*, 2394–2413. doi: 10.1080/00140139508925275
- Horswill, M. S., & McKenna, F. P. (1999). The development, validation, and application of a video-based technique for measuring an everyday risk-taking behavior: Drivers' speed choice. *Journal of Applied Psychology*, *84*, 977–985. doi: 10.1037/0021-9010.84.6.977
- Howard, R. A. (1968). The foundations of decision analysis. *IEEE Transactions on Systems Science and Cybernetics*, *4*, 211–219.
- Jackson, L., Chapman, P., & Crundall, D. (2009). What happens next? Predicting other road users' behaviour as a function of driving experience and processing time. *Ergonomics*, *52*, 154–164. doi: 10.1080/00140130802030714
- Jastrzemski, T. S., Charness, N., & Vasyukova, C. (2006). Expertise and age effects on knowledge activation in chess. *Psychology and Aging*, *21*, 401–405. doi: 10.1037/0882-7974.21.2.401
- Johnson, J. G., & Raab, M. (2003). Take the first: Option-generation and resulting

- choices. *Organizational Behavior and Human Decision Processes*, 91, 215–229.
doi: 10.1016/S0749-5978(03)00027-X
- Jones, D. G., & Endsley, M. R. (2000, October). *Can real-time probes provide a valid measure of situation awareness*. Paper presented at the Proceedings of the Human Performance, Situation Awareness and Automation: User Centred Design for the New Millennium Conference, Savannah, GA.
- Kaempf, G. L., Klein, G. A., Thordsen, M. L., & Wolf, S. (1996). Decision making in complex naval command-and-control environments. *Human Factors*, 38, 220–231. doi: 10.1177/001872089606380204
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163–182. doi: 10.1037/0033-295X.95.2.163
- Klein, G. A. (1989). Recognition-primed decisions. In W. B. Rouse (Ed.), *Advances in man-machine systems research* (Vol. 5, pp. 47–92). Greenwich, CT: JAI Press, Inc.
- Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsombok (Eds.), *Decision making in action: Models and methods* (pp. 138–147). Norwood, NJ: Ablex Publishing.
- Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 30, 576–580. doi: 10.1177/154193128603000616

- Klein, G. A., Calderwood, R., & Macgregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics*, *19*, 462–472. doi: 10.1109/21.31053
- Klein, G. A., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems*, *21*, 70–73. doi: 10.1109/MIS.2006.75
- Klein, G. A., & Peio, K. J. (1989). Use of a prediction paradigm to evaluate proficient decision making. *The American Journal of Psychology*, *102*, 321–331. doi: 10.2307/1423053
- Klein, G. A., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data-frame theory of sensemaking. In R. R. Hoffman (Ed.), *Expertise out of context: Proceeding of the sixth international conference on naturalistic decision making* (pp. 113–155). New York: Lawrence Erlbaum Associates.
- Klein, G. A., Wolf, S., Militello, L., & Zsombok, C. E. (1995). Characteristics of skilled option generation in chess. *Organizational Behavior and Human Decision Processes*, *62*, 63–69. doi: 10.1006/obhd.1995.1031
- Klinger, D. (2004). *Into the kill zone: A cop's eye view of deadly force*. San Francisco, CA: Jossey-Bass.
- Kochanek, K. D., Xu, J., Murphy, S. L., Miniño, A. M., & Kung, H.-C. (2011). *Deaths: Final data for 2009* (National Vital Statistics Reports, Volume 60, Number 3). Hyattsville, MD: National Center for Health Statistics. Retrieved from Centers for Disease Control and Prevention website: <http://www.cdc.gov/nchs/data/nvsr/>

nvstr60/nvstr60_03.pdf

- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. doi: 10.2307/2529310
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.
- Mann, D. L., Abernethy, B., Farrow, D., Davis, M., & Spratford, W. (2010). An event-related visual occlusion method for examining anticipatory skill in natural interceptive tasks. *Behavior Research Methods*, *42*, 556–562. doi: 10.3758/brm.42.2.556
- Matthews, M. D., & Beal, S. A. (2002). *Assessing situation awareness in field training exercises* (Research Report 1795). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- McGowan, A. M., & Banbury, S. (2004). Evaluating interruption-based techniques using embedded measures of driver anticipation. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness: Theory and application* (pp. 176–192). Aldershot, UK: Ashgate.
- McKenna, F. P., & Crick, J. L. (1994). *Hazard perception in drivers: A methodology for testing and training* (TRL Report 313). Crowthorne, United Kingdom: Transport Research Laboratory.
- McKenna, F. P., & Horswill, M. S. (1999). Hazard perception and its relevance for driver licensing. *Journal of the International Association of Traffic and Safety Sciences*, *23*, 26–41.

- McKenna, F. P., Horswill, M. S., & Alexander, J. L. (2006). Does anticipation training affect drivers' risk taking? *Journal of Experimental Psychology: Applied*, *12*, 1–10. doi: 10.1037/1076-898X.12.1.1
- McPherson, S. L. (1999). Expert–novice differences in performance skills and problem representations of youth and adults during tennis competition. *Research Quarterly for Exercise and Sport*, *70*, 233–251.
- McPherson, S. L., & Kernodle, M. W. (2003). Tactics, the neglected attribute of expertise: Problem representations and performance skills in tennis. In J. L. Starkes & K. A. Ericsson (Eds.), *Expert performance in sports: Advances in research on sport expertise* (pp. 137–167). Champaign, IL: Human Kinetics.
- McRobert, A. P., Ward, P., Eccles, D. W., & Williams, A. M. (2011). The effect of manipulating context-specific information on perceptual–cognitive processes during a simulated anticipation task. *British Journal of Psychology*, *102*, 519–534. doi: 10.1111/j.2044-8295.2010.02013.x
- Mehle, T. (1982). Hypothesis generation in an automobile malfunction inference task. *Acta Psychologica*, *52*, 87–106. doi: 10.1016/0001-6918(82)90028-2
- Moxley, J. H., Ericsson, K. A., Charness, N., & Krampe, R. T. (2012). The role of intuition and deliberative thinking in experts' superior tactical decision-making. *Cognition*, *124*, 72–78. doi: 10.1016/j.cognition.2012.03.005
- Mueller, S. T. (2009). A Bayesian recognitional decision model. *Journal of Cognitive Engineering and Decision Making*, *3*, 111–130. doi: 10.1518/155534309x441871
- Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive*

psychology. New York, NY: W. H. Freeman.

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

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society*, 29, 492–510. doi: 10.1017/S030500410001152X

Orasanu, J. M. (1993). Decision-making in the cockpit. In E. L. Wiener, B. G. Kanki & R. L. Helmreich (Eds.), *Cockpit resource management* (pp. 137–172). San Diego, CA: Academic Press.

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test: Different versions and factors that affect performance. *Brain and Cognition*, 28, 39–58. doi: 10.1006/breg.1995.1032

Poulton, E. (1957). On prediction in skilled movements. *Psychological Bulletin*, 54, 467–478. doi: 10.1037/h0045515

Proctor, R. W., & Van Zandt, T. (2008). *Human factors in simple and complex systems* (2nd ed.). Boca Raton, FL: CRC Press.

Raab, M., & Johnson, J. G. (2007). Expertise-based differences in search and option-generation strategies. *Journal of Experimental Psychology: Applied*, 13, 158–170. doi: 10.1037/1076-898X.13.3.158

Randel, J. M., Pugh, H. L., & Reed, S. K. (1996). Differences in expert and novice situation awareness in naturalistic decision making. *International Journal of*

- Human-Computer Studies*, 45, 579–597. doi: 10.1006/ijhc.1996.0068
- Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology*, 41, 1–48. doi: 10.1006/cogp.1999.0735
- Rumar, K. (1990). The basic driver error: Late detection. *Ergonomics*, 33, 1281–1290. doi: 10.1080/00140139008925332
- Savelsbergh, G. J. P., Onrust, M., Rouwenhorst, A., & Van Der Kamp, J. (2006). Visual search and locomotion behaviour in a four-to-four football tactical position game. *International Journal of Sport Psychology*, 37, 248–264.
- Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423.
- Shanteau, J., & Stewart, T. R. (1992). Why study expert decision making? Some historical perspectives and comments. *Organizational Behavior and Human Decision Processes*, 53, 95–106. doi: 10.1016/0749-5978(92)90057-e
- Shiffrin, R. M. (1970). Memory search. In D. A. Norman (Ed.), *Models of human memory* (pp. 375–447). New York, NY: Academic Press.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420–428. doi: 10.1037/0033-2909.86.2.420
- Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological review*, 63, 129–138. doi: 10.1037/h0042769
- Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, 41, 1–20. doi: 10.1146/annurev.ps.41.020190.000245
- Smeeton, N. J., Williams, A. M., Hodges, N. J., & Ward, P. (2005). The relative

effectiveness of various instructional approaches in developing anticipation skill.

Journal of Experimental Psychology: Applied, 11, 98–110. doi: 10.1037/1076-898X.11.2.98

Smith, K., & Hancock, P. A. (1995). Situation awareness is adaptive, externally-directed consciousness. *Human Factors*, 37, 137–148. doi: 10.1518/001872095779049444

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31, 137–149.

Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C., & Jenkins, D. P. (2005). *Human factors methods: A practical guide for engineering and design*. Aldershot, UK: Ashgate.

Stokes, A. F., Kemper, K., & Kite, K. (1997). Aeronautical decision making, cue recognition, and expertise under time pressure. In C. E. Zsombok & G. A. Klein (Eds.), *Naturalistic decision making* (pp. 183–196). Hillsdale, NJ: Lawrence Erlbaum Associates.

Strater, L. D., Endsley, M. R., Pleban, R. J., & Matthews, M. D. (2001). *Measures of platoon leader situation awareness in virtual decision-making exercises* (Research Report 1770). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Strybel, T., Minakata, K., Nguyen, J., Pierce, R., & Vu, K.-P. L. (2009). Optimizing online situation awareness probes in air traffic management tasks. In G. Salvendy & M. Smith (Eds.), *Lecture Notes in Computer Science: Vol. 5618. Human interface and the management of information: Information and interaction* (pp.

845–854). Berlin, Germany: Springer-Verlag. doi:10.1007/978-3-642-02559-4_91.

Sulistyawati, K., Wickens, C. D., & Chui, Y. P. (2011). Prediction in situation awareness: Confidence bias and underlying cognitive abilities. *International Journal of Aviation Psychology*, *21*, 153–174. doi: 10.1080/10508414.2011.556492

Suss, J., & Ward, P. (in press). Predicting the future in perceptual-motor domains: Perceptual anticipation, option generation and expertise. In J. L. Szalma, M. Scerbo, R. Parasuraman, P. A. Hancock & R. R. Hoffman (Eds.), *The Cambridge handbook of applied perception research*. New York, NY: Cambridge University Press.

Tashman, L., Harris, K. R., Ramrattan, J., Ward, P., Eccles, D. W., Ericsson, K. A., . . . Lang, L. H. (2006). Expert performance in law enforcement: Are skilled performers more effectively constraining the situation to resolve representative dynamic tasks than novices? *Human Factors and Ergonomics Society Annual Meeting Proceedings*, *50*, 1213–1217. doi: 10.1177/154193120605001201

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. *Psychological Review*, *115*, 155–185. doi: 10.1037/0033-295x.115.1.155

Tzu, S. (1983). *The art of war* (J. Clavell, Trans.). New York, NY: Delacorte Press.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Vogel, K., Kircher, A., Alm, H., & Nilsson, L. (2003). Traffic sense—which factors

- influence the skill to predict the development of traffic scenes? *Accident Analysis and Prevention*, *35*, 749–762. doi: 10.1016/S0001-4575(02)00078-7
- Ward, P. (2002). *The development of perceptual-cognitive expertise* (Unpublished doctoral dissertation). Liverpool John Moores University, United Kingdom.
- Ward, P., Ericsson, K. A., & Williams, A. M. (2012). Complex perceptual-cognitive expertise in a simulated task environment. *Journal of Cognitive Engineering and Decision Making*. Advance online publication. doi: 10.1177/1555343412461254
- Ward, P., Harris, K. R., Ericsson, K. A., Eccles, D. W., & Tashman, L. (2007). Cognitive basis for expert and superior performance in law enforcement. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*, 1884.
- Ward, P., Suss, J., & Basevitch, I. (2009). Expertise and expert performance-based training (ExPerT) in complex domains. *Technology, Instruction, Cognition and Learning*, *7*, 121–145.
- Ward, P., Suss, J., Eccles, D. W., Williams, A. M., & Harris, K. R. (2011). Skill-based differences in option generation in a complex task: A verbal protocol analysis. *Cognitive Processing*, *12*, 289–300. doi: 10.1007/s10339-011-0397-9
- Ward, P., & Williams, A. M. (2003). Perceptual and cognitive skill development in soccer: The multidimensional nature of expert performance. *Journal of Sport & Exercise Psychology*, *25*, 93–111.
- Weber, E. U., Böckenholt, U., Hilton, D. J., & Wallace, B. (1993). Determinants of diagnostic hypothesis generation: Effects of information, base rates, and experience. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, 19, 1151–1164. doi: 10.1037/0278-7393.19.5.1151

Williams, A. M., & Davids, K. (1998). Visual search strategy, selective attentive, and expertise in soccer. *Research Quarterly for Exercise and Sport*, 69, 111–128.

Williams, A. M., Davids, K., & Williams, J. G. (1999). *Visual perception and action in sport*. London, UK: Routledge.

Williams, A. M., Ford, P. R., Eccles, D. W., & Ward, P. (2011). Perceptual-cognitive expertise in sport and its acquisition: Implications for applied cognitive psychology. *Applied Cognitive Psychology*, 25, 432–442. doi: 10.1002/acp.1710

Williams, A. M., & Ward, P. (2007). Anticipation and decision making: Exploring new horizons. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (3rd ed., pp. 203–223). Hoboken, NJ: John Wiley & Sons.

Williams, A. M., Ward, P., Knowles, J. M., & Smeeton, N. J. (2002). Anticipation skill in a real-world task: Measurement, training, and transfer in tennis. *Journal of Experimental Psychology: Applied*, 8, 259–270. doi: 10.1037/1076-898X.8.4.259

Wohl, J. G. (1981). Force management decision requirements for Air Force tactical command and control. *IEEE Transactions on Systems, Man and Cybernetics*, 11, 618–639. doi: 10.1109/TSMC.1981.4308760

Wolf, S. F., Dougherty, D. P., & Kortemeyer, G. (2012). Empirical approach to interpreting card-sorting data. *Physical Review Special Topics - Physics Education Research*, 8(1), article id. 010124. doi: 10.1103/PhysRevSTPER.8.010124

Yates, J. F. (2001). “Outsider:” Impressions of naturalistic decision making. In E. Salas

& G. A. Klein (Eds.), *Linking expertise and naturalistic decision making* (pp. 9–33). Mahwah, NJ: Lawrence Erlbaum Associates.

Zimmerman, L. A. (2008). Making sense of human behavior: Explaining how police officers assess danger during traffic stops. In J. M. Schraagen, L. Militello, T. Ormerod & R. Lipshitz (Eds.), *Naturalistic decision making and macrocognition* (pp. 121–140). Aldershot, U.K.: Ashgate Publishing.

CHAPTER 9
APPENDICES

APPENDIX A

List of Scenarios Included in Studies 1 and 2

Table A1
List of Scenarios Included in Studies 1 and 2

Scenario Name	Scenario Type	Scenario Duration (s) ^a	Study 1A	Study 1B	Study 1C	Study 2
Bar	Escalating	56.81	✓	✓	✓	
Blow up	Escalating	43.71	✓	✓	✓	✓
Blue dodge ^b	Escalating	32.33	✓	✓	✓	✓
Bus lady	De-escalating	56.48	✓	✓	✓	✓
Convenience store	Escalating	39.71	✓	✓	✓	
Domestic baby	Escalating	61.33	✓	✓	✓	
Domestic punch	Escalating	29.55	✓	✓	✓	✓
Drunk	Escalating	42.26	✓	✓	✓	
Greenwood park	Escalating	19.34	×	✓	✓	
Gun grab	Escalating	45.84	✓	✓	✓	✓
Hey Mr Wilson	Escalating	67.75	×	✓	✓	✓
Hospital incident	De-escalating	50.80	✓	✓	✓	✓
Hotel	De-escalating	65.91	✓	✓	✓	
Larceny	De-escalating	87.46	✓	✓	✓	✓
Men in black	Escalating	30.71	✓	✓	✓	
Miami	De-escalating	50.99	✓	✓	✓	
Rock n roll	Escalating	34.36	✓	✓	✓	
School hostage	Escalating	40.42	✓	✓	✓	
School shooting	Escalating	49.05	✓	✓	✓	
Subway	Escalating	34.46	✓	✓	✓	✓
Suicide by cop	Escalating	93.86	✓	✓	✓	
Suicide waitress	Escalating	42.17	×	✓	✓	✓
Video store	De-escalating	38.36	✓	✓	✓	
Number of escalating scenarios:			14	17	17	7
Number of de-escalating scenarios:			6	6	6	3
Total number of scenarios:			20	23	23	10

^aIncludes introductory radio message (see Appendix E). ^bThis trial was used for the practice trial.

APPENDIX B

Exemplar Scenarios

Blow Up

Radio message/context. “You and your partner are responding to a call describing a disturbance outside of a school. A boy sitting on the grass directly in front of you has been acting suspiciously.”

Scenario description. The participant arrives on scene with another officer. The fellow officer escorts students away from the suspect and asks the participant to take care of the suspect. As the participant approaches the suspect, the boy gets up and becomes verbally aggressive. The boy is wearing an untucked, partially buttoned shirt. A wire can be seen hanging down under the shirt. The boy opens his shirt to expose an improvised explosive device around his waist. A school bus approaches in the background. The boy removes the actuator from his waistline and holds it in his hand. He then gets flustered and drops the actuator on the ground. He bends down to pick it back up and then detonates the device just as the school bus pulls up next to him.

Convenience Store

Radio message/context. “While on patrol, you stop at the neighborhood convenience store for a cup of coffee. You are getting out of your car and going into the store.”

Scenario description. The participant enters the store. The store manager greets him and nervously invites him to stay and have a cup of coffee. The camera pans to allow the participant to look around the store (and away from the store manager). While

looking around, the sound of a round being chambered into a firearm can be heard. The camera pans back around to show the perpetrator holding the store manager hostage. After a few seconds, the perpetrator surrenders, puts his weapon down, and releases the store manager. Immediately afterwards, the perpetrator then draws a backup weapon and shoots in the direction of the participant.

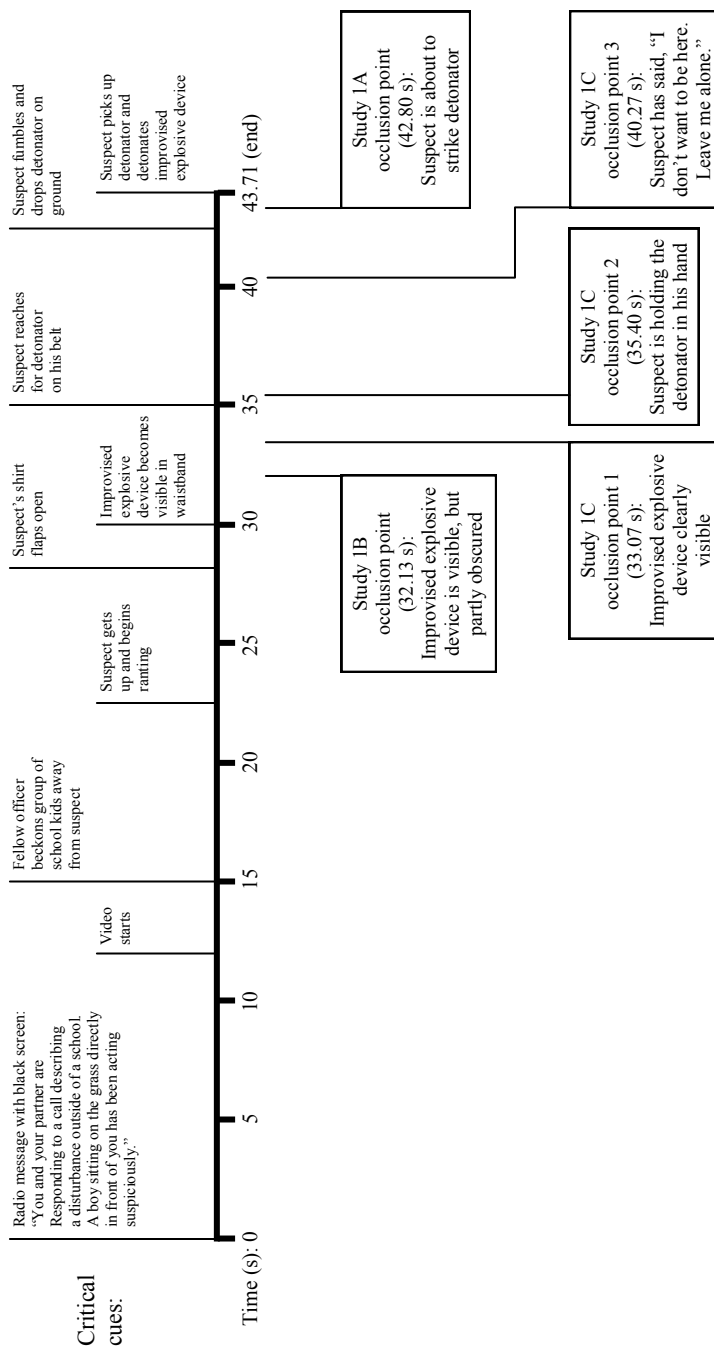
School Hostage

Radio message/context. “You and your partner are responding to a call describing a possible hostage situation at a school. You are in the passenger seat and your partner is driving to the scene.”

Scenario description. The officers drive into the school’s service (i.e., back) entrance. After entering this area, just as they stop and begin to get out of the car vehicle, a male suspect armed with an assault rifle becomes visible, close to the rear entrance to the school. The male immediately turns to run back into the school and appears surprised to see the police. The suspect opens the rear door and enters the school with the assault rifle.

APPENDIX C

Timeline of Blow Up Scenario with Description of Critical Cues and Timing of Occlusion Points



APPENDIX D

List of Occlusion Point Times (in seconds) for Studies 1A–C

Table D1

List of Occlusion Point Times (in seconds) for Studies 1A–C

Scenario	Study 1A	Study 1B	Study 1C		
	Occlusion point	Occlusion point	Occlusion point		
	1	1	1	2	3
Bar	55.47	54.93	54.00	54.87	55.40
Blow up	42.80	32.13	33.07	35.40	40.27
Blue dodge ^a	29.60	29.60	29.60	n/a	n/a
Bus lady	54.13	54.20	48.33	53.73	54.20
Convenience store	38.53	38.47	38.27	38.47	38.67
Domestic baby	59.13	59.27	57.60	59.13	59.20
Domestic punch	27.73	27.93	27.40	27.93	28.27
Drunk	40.40	40.40	40.93	41.40	41.47
Greenwood park	–	15.87	13.53	13.93	15.93
Gun grab	43.47	39.67	39.73	40.73	41.60
Hey Mr Wilson	–	66.73	62.89	63.80	66.20
Hospital incident	38.27	38.27	38.33	38.60	39.80
Hotel	54.07	54.07	52.27	54.47	55.20
Larceny	61.93	59.67	52.20	53.67	59.47
Men in black	29.53	29.73	28.60	29.00	29.73
Miami	40.73	40.67	38.00	39.87	40.67
Rock n roll	28.40	28.93	28.93	29.33	29.87
School hostage	37.13	37.13	36.87	36.93	37.00
School shooting	47.53	47.47	47.40	47.47	47.53
Subway	31.60	31.60	31.60	31.67	31.73
Suicide by cop	74.33	74.20	73.00	74.07	74.13
Suicide waitress	–	41.13	41.20	41.27	41.33
Video store	33.20	34.20	33.93	34.00	34.07

Note. Dashes indicate scenarios that were not included in Study 1A. n/a = not applicable.

^aThis trial was used as the practice trial in Studies 1 and 2; the same occlusion point was used across all studies.

APPENDIX E

Radio Message/Contextual Introductions to the Video Scenarios

Table E1
Radio Message/Contextual Introductions to the Video Scenarios

Scenario	Radio message/context
Bar	“You have been called to a bar to deal with a patron.”
Blow up	“You and your partner are responding to a call describing a disturbance outside of a school. A boy sitting on the grass directly in front of you has been acting suspiciously.”
Blue dodge	“You make a traffic stop. You are walking to the car that you stopped.”
Bus lady	“You have been sent to a bus stop to deal with a passenger.”
Convenience store	“While on patrol, you stop at the neighborhood convenience store for a cup of coffee. You are getting out of your car and going into the store.”
Domestic baby	“You are answering a call to a domestic dispute and are about to enter the house.”
Domestic punch	“You are answering a call to a domestic dispute and are about to enter the house.”
Drunk	“You are standing in a parking lot. There is man standing directly in front of you, acting suspiciously.”
Greenwood park	“An undercover officer has called for backup to Greenwood Park.”
Gun grab	“You have been dispatched as a backup unit to a local hotel, where a suicidal man is threatening to jump from the building.”
Hey Mr Wilson	“You are answering a call to a domestic dispute.”
Hospital incident	“You are responding to the call of a disturbance at a local hospital. You are standing in the hospital waiting room.”
Hotel	“You and your partner are investigating a possible suspect with two arrest warrants. The suspect has long hair and is wearing boots and jeans. You are standing in an elevator, looking at your partner.”
Larceny	“You are responding to a call of possible larceny. You are walking towards a parked train.”

(Table E1 continues)

(Table E1 continued)

Scenario	Radio message/context
Men in black	“You have made a traffic stop. You are standing outside of your car with the suspect.”
Miami	“You are responding to a call of a suspicious car parked outside of a building. You are standing outside of the vehicle on the driver’s side.”
Rock n roll	“You have been dispatched to an industrial estate, where a male was reported to be loitering.”
School hostage	“You and your partner are responding to a call describing a possible hostage situation at a school. You are in the passenger seat and your partner is driving to the scene.”
School shooting	“You and your partner are responding to a call describing a disturbance at a school. You are running down the hallway inside the school.”
Subway	“You have been called to a reported disturbance at a subway station.”
Suicide by cop	“You are responding to a call describing a possible suicide attempt in progress. You are standing in a hallway. Directly in front of you is a woman.”
Suicide waitress	“You are responding to a report of a suicidal female at a restaurant.”
Video store	“You are responding to a call of a disturbance at a local video store.”

APPENDIX F

G*Power Power Analysis Output

[1] -- Monday, December 03, 2012 -- 16:34:49

z tests – Proportions: Difference between two independent proportions

Analysis: A priori: Computer required sample size

Input:	Tails(s)	=	One
	Effect size h	=	0.75
	Proportion p1	=	.51
	α err prob	=	0.05
	Power (1- β err prob)	=	0.80
	Allocation ratio N2/N1	=	1
Output:	Critical z	=	1.6448536
	Sample size group 1	=	23
	Sample size group 2	=	23
	Total sample size	=	46
	Actual power	=	0.8078179
	Proportion p2	=	0.8480681

APPENDIX G

Thought Statement Types

Table G1
Thought Statement Types Used to Code Retrospective Verbal Reports

Thought statement type	Description
Monitor	Heeded information that was present in the current environment (e.g., “The suspect was standing in front of me”)
Infer	Information that is inferred about some aspect of the current scenario, but where that information is not present in the current environment (excluding evaluative inferences; e.g., “If my partner doesn’t turn the siren off, the hostage takers will hear us approaching”)
Evaluate	A relative value-based inference about some aspect of the past, current or future environment but where the inferred value or information is not present in the current environment (e.g., “The boy appeared to be depressed”)
Predict	An anticipated future situational event (e.g., “I think he’s going to go back in the school”)
Act	A verbal or physical action and/or execution of a decision (e.g., “I told him to release the store manager”)