2017

Sampling expertise: Incorporating goal establishment and goal enactment into theories of expertise to improve measures of performance

Frank Eric Robinson
Wright State University

SAMPLING EXPERTISE: INCORPORATING GOAL ESTABLISHMENT AND GOAL ENACTMENT INTO THEORIES OF EXPERTISE TO IMPROVE MEASURES OF PERFORMANCE

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

By

FRANK ERIC ROBINSON
B.A., Trinity University, 2007
M.S., Wright State University, 2011

2017
Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

APRIL 18, 2017

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Frank Eric Robinson ENTITLED Sampling Expertise: Incorporating Goal Establishment and Goal Enactment into Theories of Expertise to Improve Measures of Performance BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Valerie Shalin, Ph.D.
Dissertation Director

Scott Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Final Examination

John Flach, Ph.D.

Joe Houpt, Ph.D.

David LaHuis, Ph.D.

Robert E.W. Fyffe, Ph.D.
Vice President for Research and
Dean, Graduate School

Abstract

Robinson, Frank Eric. Ph.D., Department of Psychology, Wright State University, 2017. Sampling Expertise: Incorporating Goal Establishment and Goal Enactment into Theories of Expertise to Improve Measures of Performance

Task-specific performance measures informed by incomplete theories of expertise do not capture the full range of domain-relevant behaviors, threatening content validity. Surgery is a particularly good example of a domain that has neglected cognitive accounts of performance in favor of task-specific measures of technical skill and experience-based definitions of expertise. Likewise, cognitive accounts of performance tend to neglect skilled performance, including the interaction between automaticity and cognitive control. The present study merges cognition and psychometrics in the context of a surgical task. I analyzed archival surgical performance data from a study of surgical training, including video of human cadaver procedures, think-aloud, self-ratings, and performance evaluations. This rich data set provided a unique opportunity to address both theoretical and methodological issues within expertise research, such as the ability of generalizable constructs to account for task-specific performance measures, the cognitive penetrability of skilled performance, the contribution of experience to the development of expertise, and the impact of evaluator cognition on performance ratings. My analyses indicate that general constructs related to goal establishment and goal enactment can account for task-specific performance metrics, highlighting the cognitive penetrability of skilled performance in the process. My analyses also call into question the necessity of experience in the development of expertise, and illustrate the influence of the evaluators on performance ratings. Accounting for these elements will strengthen theories of performance and subsequently help promote measures of performance that will generalize within a domain rather than emphasize any particular task.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ANALYSES

ACKNOWLEDGEMENTS

First, I would like to express by sincere gratitude to the entire research team at the University of Maryland for allowing me to use their data for my dissertation. In particular, I would like to thank Colin Mackenzie, Evan Garofalo, and Kristy Pugh for going far beyond any reasonable expectation to provide insight and answer questions along the way.

I would also like to thank my advisor, Valerie Shalin, for her mentorship. She has been a constant (and hopefully continuing) source of wisdom, insight, and guidance, and I am much better for it.

In addition, I thank John Flach, Joe Houpt, and David LaHuis for serving on my dissertation committee. Their advice regarding my analyses and comments on the draft of this paper greatly strengthened the dissertation.

Finally, I would like to thank my family. I thank my wife, Jasmine Duran, for her unwavering love and support. She has been extremely understanding and a great ear for ideas and frustrations. I hope to return the favor when it comes time for her dissertation. I also thank my new daughter, Alessa, whose impending arrival gave me a final kick in the pants to finish.

Chapter 1 – Introduction

Performance measures guided by theories of expertise have the potential to be more principled and generalizable than performance measures guided by task-specific analyses. For example, Patel & Groen (1986) propose that expert physicians rely on forward reasoning based on evidence to arrive at medical diagnoses, whereas less proficient doctors reason backwards based on hypotheses. A performance measure designed to examine these thought processes may distinguish between experts and novices in a variety of contexts. Unfortunately, Patel's and other laboratory-based theories of expertise rely on artificially parsed problems and neglect aspects of expertise such as problem identification and enacting solutions in context.

Task analyses (driven by varied conceptualizations of expertise) form the foundation for methods of performance evaluation – incomplete task analyses result in metrics that fail to capture the full range of behaviors that characterize expertise, threatening content validity. Task-specific performance evaluations have face validity, but the absence of a theoretical foundation limits the generalizability of such approaches to the family of related activities (Clancey). On the other hand, the current theoretical foundations require expansion to guide the systematic sampling of domain-relevant behaviors.

The present study addresses these issues via archival surgical performance data from a study of surgical training. The available data included video of human cadaver procedures, performance ratings, self-evaluations, and think-aloud data from the surgeons.

This rich data set provided a unique opportunity to address both theoretical and methodological issues within expertise research. Surgery is a particularly good example of a domain that has neglected cognitive accounts of performance in favor of task-specific measures of technical skill. Likewise, cognitive accounts of performance tend to neglect skilled performance in favor of knowledge-oriented measurement. Merging cognition and psychometrics in the context of a surgical task more fully captures expert behavior and provides insight into the balance between cognitive control and automaticity in expert performance. Incorporating these elements will strengthen theories of performance and subsequently promote domain-general measures of performance that extend beyond any particular task.

The remainder of the introduction addresses past research on expertise while advocating for the role of cognitive penetrability in expert performance. I offer a critique of cognitive models, arguing in favor of the inclusion of constructs such as goal establishment and goal enactment. I describe these constructs along with how each is affected by deliberate processes and how the two constructs interact with one another. I also describe the medical view of expertise, and the resulting emphasis on criterion-oriented performance measures. Finally, I address the difficulty in capturing expert performance using such measures before previewing the rest of the paper.

1.1 The Nature of Expertise

By itself, expert behavior is difficult to identify. Such behavior only becomes apparent when disrupted or by examination during development in less skilled individuals. Cognitive conceptualizations of expertise based on expert-novice comparisons are largely grounded in the organization of knowledge, identified via comparisons between experts and novices. Cognitive research has identified three characteristics that allow experts to behave adaptively in various contexts: declarative knowledge, procedural knowledge, and contextual flexibility (Dreyfus & Dreyfus, 1986). Experts' knowledge base and ability to make associative connections in the world are developed over time based on deliberate practice intended to improve skill and knowledge (Ericsson, Krampe, & Tesch-Romer, 1993). Based on more complex and structured associative networks, experts think at a higher level of abstraction than novices, viewing problems at a deeper, more principled level (Chi, Feltovich, & Glaser, 1981; Tanaka & Taylor, 1991). Experts are able to use forward reasoning, utilizing data to drive hypotheses and decision making (Patel & Groen, 1986). Expert behavior is largely automatic, freeing resources to allow better strategic self-monitoring compared to novices (Beilock et al., 2004; MacIntyre et al., 2014; McPherson, 2000).

Based on these abilities, experts approach problems in a fundamentally different way than novices. Good reasoning requires the ability to perceive the key functional relationships of a problem to arrive at a proper understanding of the relevant constraints and possibilities of a situation (Duncker, 1926; Duncker, 1945). Experts are able to use various types of processes based on situational demands and constraints: skill-based, rule-based, and knowledge-based behavior (Rassmusen, Pejtersen, & Goodstein, 1994). Rasmussen's behavioral distinction posits that people interact with the world differently depending on whether information is linked with the world directly (skill-based), by consistent association (rule-based), or by convention (knowledge-based). Skill-based behavior does not require deliberation, while knowledge-based behavior requires effortful reasoning. These different levels of processing, based upon the expert's familiarity with the domain in question and the nature of the link between information and the world, allow the expert to utilize skill-based reasoning for routine situations but still utilize more effortful representational reasoning when necessary.

In a similar vein, Klein's (1989) recognition primed decision making (RPD) model of expert decision making posits that experts do not deliberate. Instead, they recognize situations and rely on associative processes to arrive at a solution. Mental simulation determines the suitability of an associatively generated solution. If this fails, the expert deliberates more thoughtfully and engages in more traditional problem solving.

1.1.1 Problem solving in the world demands cognitive penetrability.

Experts can move adaptively through a decision ladder, taking shortcuts (skill- or rule-based behavior) when possible but deliberating (knowledge-based behavior) when required (Vicente, 1999). Experts rely on both associative and deliberative processes to generate solutions to problems in the world. However, the expertise literature largely treats automatic and deliberate processes as separate entities. For example, Rasmussen's (1994) SRK framework and Klein's (1989) RPD model each only involve deliberative thought processes in unique or demanding situations, or upon failure of more associative processes. Left unaddressed is the extent to which associative/skill-based and deliberate/knowledge-based reasoning interact with one another, particularly the influence of higher-level processes on skilled behavior.

Higher-level processes can influence behaviors in two ways: the allocation of attention, and decisions in recognizing and identifying patterns (Pylyshyn, 1999). No two situations are alike – solving problems in the world demands adapting to particular contexts in order to respect sometimes-competing situational constraints and overcome novel situations. Conventional wisdom asserts that skills become more automatic with practice, precluding conscious control and even implying that conscious awareness can be harmful to performance (Baumeister, 1984; Lewis & Linder, 1997). However, more recent work indicates that deliberate cognitive control strategies can help maintain and even improve performance during stressful situations (Balk et al., 2013; Toner & Moran, 2014). Implicit behaviors are not isolated from explicit processes, and people are able to reflect on and control the outcomes of implicit processes.

A key issue in appreciating the role of knowledge in expertise concerns the ability to influence action with goals and belief, an issue known as cognitive penetrability. If the output changes based on the agent's knowledge, then that system is cognitively penetrable (Pylyshyn, 1999). Penetrability can mean conscious processes influencing or overriding unconscious outputs. The cognitive penetrability of processes related to surgery may help shed light on how experts balance automatic (skilled) and deliberate (representational) decision making.

Though the cognitive literature has distinguished between automatic and deliberative processing for some time, the interaction between the two is a relatively recent topic of research (Pennycook, Fugelsang, & Koehler, 2015). An emerging viewpoint is that automatic processes are triggered by the environment, and monitored by deliberative processes. In cases of agreement, the two modes work synergistically. If conflict is detected between the two modes of operating, more effortful thinking is engaged to override heuristic processes (Ferreira et al., 2016; Pennycook, Fugelsang, & Koehler, 2015). Both top-down (goal based) and bottom-up (conflict based) processes can engage analytic reasoning (Ferreira et al., 2016). These studies have only applied to decision making processes, however, rather than the skilled behavior characteristic of many domains of expertise.

Experts can evaluate and double check the outputs of their heuristic processing to determine if that truly is the best course of action. Even highly automatic behaviors such as reading can be subject to deliberate control. Increased reading skill is associated with reduced interference on font color naming in the Stroop task, indicating better ability to inhibit the supposedly automatic reading response (Protopapas, Archonti, & Skaloumbakas, 2007). Both human Stroop data and cognitive (ACT-R) modeling of the Stroop task indicate that some degree of inhibition is necessary to perform the task. Although modeling data indicates that such inhibition may be an emergent property of more fundamental mechanisms, the *function* of inhibition is still important for skilled performance on the task (Juvina, 2011).

1.1.2 Declarative knowledge only gets us so far in test development.

Knowledge is necessary but not sufficient for skilled performance. Typical paper and pencil job knowledge tests correlate with performance, but not highly enough to support using knowledge tests as a substitute for hands-on assessments (DuBois & Shalin, 1995). Such tests tap declarative knowledge, which is distinct from performance; having knowledge does not guarantee that it will be used correctly (Hammond & Summers, 1972). Knowledge must be linked with context to allow performance in real-world tasks (Dunphy & Williamson, 2004). The process of execution must be captured in addition to domain knowledge.

As suggested above, experts must be able to adapt to unique situations. Conventional paper-and-pencil measures of knowledge fail to account for how knowledge is used for performance (DuBois & Shalin, 1995). For example, one of the key differences between residents and attending physicians seems to be in knowing how to do something rather than knowing what should be done (Abernethy et al., 2008).

Expert surgeons do not necessarily have more declarative domain knowledge than advanced trainees. Rather, the experts likely have better control; they have a better awareness of their knowledge, are better able to apply procedural knowledge, and are better able use their knowledge to deal with uncertainty. Focusing on what should be done rather than how to do it has limited laboratory (paper and pencil) methods for revealing expertise.

1.1.3 Process-based accounts are better suited to capturing expertise.

Ericsson and Lehmann (1996) advocate for identifying expertise based on consistently superior performance rather than other criteria such as experience or peer nomination, but identifying such performance objectively is not always straightforward. Two different approaches regarding the definition of expertise are evident in the cognitive literature: criterion-based approaches (e.g., the expert performance approach) or relative approaches (e.g., the performance based approach). Criterion-based approaches define expertise based on outcome, arguing that expertise should be identified and operationalized based on reaching a predefined level of exceptional performance (Ericsson & Lehmann, 1996). Like the criterion-based approach, the relative approach requires objective measurement of performance. However, the relative approach argues that experts are better defined as superior relative to others within a domain rather than with respect to a given threshold (Weiss & Shanteau, 2014a).

The relative approach favors process rather than outcome in order to accommodate domains with outcomes that are beyond the expert's control, subjective, or otherwise difficult to measure (Weiss & Shanteau, 2014a; 2014b). Performance becomes defined as a behavior rather than as a quality metric (Weiss & Shanteau, 2014b). Because such definitions of expertise rely on behavior rather than outcomes, they are better able to accommodate aspects of expertise related to planning, monitoring, and contextual adaptation (i.e., goal establishment and goal enactment). Constructs such as these are typically excluded from conventional problem solving research.

1.2 Architectural Perspectives on Expertise

The above cognitively-oriented conceptualizations of expertise often draw on constructs that arise from the specification of general cognitive architectures, typically concerning the organization and retrieval of knowledge (typically for decision making). Such a notion forms the basis of cognitive architectures such as ACT-R, which then drives theories of expertise and inform the dimensions of skill identified via task analyses. ACT-R assumes two types of knowledge: declarative and procedural (Anderson, Matessa, & Lebiere, 1997). ACT-R instantiates goals as if-then production rules, but these syntactically oriented, descriptive accounts of procedure are too brittle to capture the essence of expertise. For example, ACT-R models automaticity by combining production rules to form procedures, increasing efficiency but suggesting difficulty in inhibition and the absence of penetrability (Anderson, 1992). Procedural knowledge becomes uninspectable.

ACT-R relies on modules for various cognitive functions (e.g., a vision module, declarative module, goal module, etc.), connected via a procedural module that fires task-specific production rules determined by the researcher (Nyamsuren & Taatgen, 2013). The architecture's goals are pre-determined. ACT-R was originally intended to model high level cognition. As a result, it largely ignored interactions with the world via vision or attention (Anderson, Matessa, & Lebiere, 1997). Visual modules were added later to address this issue, but the inputs are still mediated via a buffer rather than used directly by the modules in ACT-R (Anderson, Matessa, & Lebiere, 1997). Nonrepresentational skilled behavior is therefore difficult to capture appropriately. Further, ACT-R's vision is limited in how it can explore the world. ACT-R was designed in an assumed environment that excludes large spaces beyond the visual field. Only recently have modifications been made to allow for things like head movement (Oh, Jo, & Myunh, 2014), to say nothing of tactile or other interaction with the environment.

Models such as these may not reflect expert cognition. Anderson's perspective neglects the role of cognitive control in automaticity and fails to allow for reflection or control once a procedure is activated. Procedural knowledge instantiated in this way addresses states of the world that become true upon completion of production rules. However, execution in the context of a continuously variable world is poorly addressed. Experts demonstrate more flexibility than these models anticipate – the ability to form new goals, monitor outcomes, and adjust their behaviors to fit unique circumstances in real time (Smith, Greeno & Vitolo, 1989; Greeno, Riley and Gelman, 1984). These considerations suggest that conventional architecturally-motivated distinctions, such as declarative and procedural knowledge offer limited guidance in the specification of expertise.

1.2.1 Looking inside procedural knowledge: Goal establishment and goal enactment.

The predominant cognitive conceptualization of expertise as grounded in the organization of knowledge has largely constrained the study of expertise to tabletop measures, focused on how experts reason about a given problem and information set. At the same time, the relative ease of access to knowledge has further encouraged its basis for theories and models of expertise. Laboratory tasks of expert reasoning tend to focus on comparison of alternatives and normative models of reasoning based on represented problems. These problems neglect the perceptual, attentional, and memory processes involved. Such problems are bounded, as opposed to the dynamic real world (Flach et al., in press). A more principled heuristic for sampling behaviors relevant to expertise is necessary to capture a wider range of behaviors relevant to skilled performance in . We must capture how experts form and enact goals in the context of a real work environment rather than the parsed and limited world of a laboratory.

Many laboratory paradigms of expertise study problem solving by giving the subject a pre-defined diagnostic problem to reason about (e.g., Patel). Such studies examine how knowledge is organized and used to reason about represented problems, but do not account for how experts take action within the constraints of the physical world. The world is naturally unparsed - experts must make sense of it in order to act. Experts must also be able to parse the world in a way that supports action; providing experts with an artificially parsed problem set (i.e., providing all possible diagnostic information with no way to filter it appropriately) can impair decision making (Kulatunga-Moruzzi, Brooks, & Norman, 2004). The typical cognitive experimental paradigm neglects how experts first identify problems in an open (i.e., not mediated by the representations used in the lab) world and then enact a solution in the context of the work environment. Applying knowledge to solve a problem in the real world requires two inter-related but conceptually separable components of expertise: goal establishment and goal enactment (Robinson, 2011).

Although cognitive architectures merge these constructs in production rules, goal establishment and goal enactment behaviors are separate psychological constructs; each responds differently to environmental stressors and demonstrates a different developmental trajectory in the ability to fit behaviors to contexts (Robinson, 2011). Goal establishment behaviors enable the expert to identify problems that require attention and form a preliminary plan of action. Goal enactment behaviors help complete the planned course of action while confirming that the plan is appropriate. The distinction is essentially a question of "what" vs. "how", similar to the distinction between function-oriented task models and action-oriented activity/process models (Clancey). While the former models are more useful for describing the interactions among elements of complex systems, both must be utilized in order to understand the system as a whole.

Goal establishment can be critical to successfully solving problems, as it shapes future actions during the problem solving process. In fact, knowledge of goals correlates with performance more strongly than declarative knowledge (DuBois & Shalin, 1995). Goals arise from an effort to remove threats to a desired state of the world (Lippa et al, 2016). Successful goal establishment depends on first realizing that such a threat has occurred and then determining the specific nature of that threat. The parameters and constraints of a situation specify what may be done to address such threats, but may not specify the problem entirely (Simon, 1973). Goal establishment is the process of identifying the relationship between the current state, the desired end state, and the possibilities offered by the world, and translating those relationships into coherent goals and plans.

Once a threat to a desired state of the world has been identified, one must begin to enact a solution to mitigate the threat - what I term goal enactment. Measures of expertise using pre-defined problems ask experts to name the solution, but a solution is only effective if implemented. The ability to identify and remove threats to desired system states develops with expertise and is strengthened by the incorporation of higher level processes.

1.2.1.1 Expertise in goal establishment and goal enactment.

Experts and novices adapt their goal establishment and goal enactment behaviors differently in response to changing circumstances (Robinson, 2011). The nature of error tends to vary with experience (Weaver, Newman-Toker, & Rosen, 2012), indicating that skill in goal establishment and/or goal enactment changes with increasing skill. One strategy experts can use to approach novel or complex problems within their areas of expertise is to decompose the large problem into a series of gradually more specific subproblems (Schraagen, 1993; Simon, 1973).

Perception and action are intimately coupled. Action in the world can serve as a useful source of information (Flach & Warren, 1995). Experts learn to make finer distinctions among the perceptual information available and can respond to features of environmental stimuli that novices do not detect (Gibson & Gibson, 1955). With increasing knowledge and expertise, surgeons are better able to interpret their interactions with the world. Novices or less-skilled people rely heavily on environmental feedback from information seeking behavior (Sadideen et al., 2013). If uncertain about whether the proper artery has been identified, a surgeon can constrict the structure in question and observe whether bleeding stops. However, this strategy poses risks. Long-term damage that is not immediately apparent may result if a nerve is constricted instead of the artery. Over time, however, people learn to identify positive indicators of performance without negative indicators from information seeking interventions (Schmidt & White, 1972). This self-monitoring ability allows experts to avoid negative consequences as they act in the world; they can monitor and correct performance in real time and do not depend on a negative outcome to identify potential mistakes.

1.2.1.2 Deliberate processes in goal establishment.

Goal establishment requires attunement to the relevant cues of the world. Certainly, people can direct their attention deliberately in order to assess different sources of information (Posner, 1980; Wolfe, 1994). However, relevant cues in the world may change based on goals and situational constraints (Hosking, Davey, & Kaiser, 2013). People must allocate attention in a way that matches task demands (Smith et al., 2001; Stanard et al., 2012). Experts are more attuned to the constraints of the environment, allowing them to direct attention to relevant stimuli (DeGroot, 1965; Vicente & Wang, 1998) and facilitating goal establishment.

The role of deliberation in goal establishment has implications for the assessment of expertise. The ability to gather information deliberately may require expertise. Expert emergency physicians are able to adapt their goal establishment behaviors to changing conditions, but nonexperts are not (Robinson, 2011). When a student first enters a domain, important stimuli tend to intrude fortuitously on the student rather than being controlled by the student. With practice, students learn to incorporate these elements into their goal structure, new aspects of the environment become relevant, and students begin to integrate them into their cognitive processes (Simon, 1967). In other words, people learn to look for important cues. Knowing how to seek out relevant information can help experts not only identify goals, but monitor progress to identify threats to the attainment of those goals in real time.

1.2.1.3 Deliberate processes in goal enactment.

Supposedly automated behavior may be cognitively penetrable. Intention, belief, metacognition and self-monitoring, and directing attention all play a part. The processes of intellectual reasoning and motor control are linked. For instance, framing effects apply to motor behavior - psychomotor tasks are executed differently depending on how a task is framed (Huhn, Potts, & Rosenbaum, 2016). I do not argue that all procedural skill requires conscious control; rather I argue that conscious processes can facilitate execution in context via planning and monitoring. Automaticity may actually facilitate deliberation by freeing cognitive resources (Dunphy & Williamson, 2004).

1.2.2 The interaction between goal establishment and goal enactment.

Despite representing unique constructs, goal establishment and goal enactment are tightly coupled. Solving complex problems in the world over time most often requires a process akin to adaptive control, wherein one continually adjusts behavior to satisfy functional goals (Flach & Voorhorst, 2017). In adaptive control, two intimately coupled control loops work together. An inner perception-action loop executes behavior, while an outer supervisory loop monitors for anomalies and violated expectations as a result of the activities of the inner loop. Should the outer loop detect a problem, it adjusts the operation of the inner loop to bring the system back in line with the desired state (Flach et al., in press). Goal establishment (here analogous to the outer control loop) and goal enactment (the inner loop) work together in a similar way - this interaction must be incorporated into theories of expertise in order to sample behaviors properly and drive improvements to performance measurement.

Though conceptually distinct, the classification of an individual observed behavior as goal establishment or goal enactment is best viewed as functional. Goal establishment links a problematic initial state to a preferable final state based on the cues and constraints of the world. Goal enactment effects the desired change, thereby creating a new state of the world and the context in which later goal establishment occurs. The real time coupling of goal establishment and goal enactment helps to ensure a match between solution and problem. People must be able to respond to changes in the world that demand adjustment to the goal hierarchy. This requires a degree of continuous processing to detect when interruption of ongoing processes is necessary (Simon, 1967). Similar to the inner and outer control loops described above, monitoring is required to ensure that goal enactment processes are appropriate during task execution.

Goal establishment and goal enactment interact and may share many fundamental cognitive processes. Nonetheless, the distinction provides an essential guide to domain sampling, as current outcome-oriented performance evaluations are unlikely to reflect this functionality. Criterion-based measures aggregate all aspects of performance into a single score, neglecting *how* the expert arrived at the result. As a result such measures may miss important distinctions in the establishment or enactment processes that would distinguish between apparently equally skilled individuals.

1.2.2.1 Goal enactment informs goal establishment.

Action in the world generates new cues and problem states that help specify goal establishment behaviors. In this way, goal establishment and goal enactment are iterative as one process informs the other. One illustration of the iteration occurs in the case of error, where goal enactment informs goal establishment. Consider a surgeon who gets "lost" within the patient. The surgeon may realize something is wrong and begin to assess the general nature of the problem (e.g., "I may not be handling the artery."). Eventually, the surgeon is likely to realize the exact nature of the problem (e.g., "This is the axillary nerve, but I need the axillary artery."). Here, engagement with the world promotes the interaction between establishment and enactment. Handling fibrous nerves for example feels different from handling arteries or veins, which also have a differential feel. The surgeon will use that knowledge to modify enactment with a new subgoal (e.g., "If this is the nerve, I need to look *deeper* for the artery."). This episode also illustrates the role of knowledge to guide enactment, by relying on anatomical heuristics regarding the relationship between arteries, veins and nerves. However, this heuristic only works with awareness of one's location in the body.

Goal enactment may also facilitate future goal establishment by generating outcomes that structure the world to make future problems and possible solutions more evident. For example, cleaner cutting technique may facilitate navigation through the body or the identification of anatomical structures. New problems can be detected during the process of solving the original problem - planning and action can inform one another (Kirsh & Maglio, 1994). More generally, action can promote an understanding of the world (Flach & Warren, 1995).

1.2.2.2 Goal establishment informs goal enactment.

Goal establishment also determines goal enactment, in part reducing replanning. Experts set themselves up for success by shaping their environment based on their assessment of the problem – subsequently positioning their own body or the patient to facilitate work towards the identified goal. Inadequate goal establishment will lead to suboptimal setup for goal enactment. Expert surgeons also anticipate what instruments they will need in advance, reducing the need to search for instruments and allowing the expert to maintain focus on the surgical field (Tien et al., 2015), facilitating monitoring of the procedure. Surgeons must also respect situational constraints including hospital policy, equipment/staff availability, and variable patient anatomy. Establishment processes may occur in the midst of enactment processes in order to ensure such constraints are respected.

In sum, establishment and enactment interact to influence one another and facilitate goal attainment in the world. Performance measures must sample such behaviors in order to better capture the full range of behaviors that identify expertise.

1.3 Testing Implications

Establishment and enactment contribute to expert behavior in the world, but are excluded from many analyses of domain behaviors that drive performance measures. Though evolving, the medical perspective on expertise is an example of a domain that neglects these issues in favor of technically-oriented indicators of skill. As a result, current performance measures have difficulty distinguishing among experts.

1.3.1 Medicine takes a narrow perspective on expertise.

Surgical skill has traditionally been viewed as based on innate traits; only more recently have other factors such as criterion-referenced skill and nontechnical factors been incorporated (Alderson, 2010). Expertise in surgery is poorly defined. Surgeons' and educators' definition may only reflect basic competency rather than expertise (Gelinas-Phaneuf & Del Maestro, 2013). Surgical performance measures have reflected this limitation. Measures of surgical performance tend to favor criterion-based approaches rather than process-based approaches, and fare poorly at evaluating experts. A more holistic viewpoint incorporating goal establishment and goal enactment will help increase the validity of surgical performance measures.

Most medically-oriented models of surgical expertise focus on aspects of performance related to task execution such as perception, cognition, action production, attention, and feedback utilization (Abernethy et al., 2008). Moreover, the specificity of the resulting items limit evaluations to individual tasks within a domain. By focusing on task-specific behavior and overlooking functionally oriented constructs such as the effect of context on behavior, planning and decision making, medical evaluations of expertise risk overlooking key aspects of performance. A more complete view of expertise will help guide domain sampling to better capture expertise. By capturing more general markers of expert performance, we can measure skill within a domain rather than a particular procedure to better gauge overall competence.

1.3.2 Current surgical performance measures do not capture expertise.

Most of the methods used to measure medical performance do not meet the criteria for effective assessment and scores are often based on subjective judgment rather than objective measurement (Mitchell et al., 2014). Current metrics can distinguish among trainees, but hit a ceiling when evaluating expert surgeons (Ahmed et al., 2011). Most tools are not designed with experts in mind; they focus on core competencies such as patient care, medical knowledge, professionalism (Ahmed et al., 2011). Several types of measurement tools are commonly used, including procedural logs, written or oral exams, checklists, global rating scales, procedure time, error rates, motion analysis, self-report, video analysis, and ratings of nontechnical or sociocultural factors (Mitchell et al., 2014; van Hove et al., 2010). Many of these are simply formats populated subjectively with no principled guidance for what may be included in them. Content-oriented measures such as motion analysis, however, are becoming increasingly popular.

1.3.3 The surgical perspective on expertise favors criterion-based performance measures.

There is a trend towards using technical competence as a measure of trainee readiness rather than number of procedures performed, but objective measures of competence are not well reported or studied (Neequaye et al., 2007). Surgical performance measurement is too task-specific; the surgical domain lacks validated models of expertise to guide behavioral sampling.

The medical community is shifting to more criterion-based measures of performance (Mitchell et al., 2014), but process is typically neglected. Training evaluations often use measurable patient outcomes such as complications or mortality. Process measures such as blood loss or operative time are sometimes included (Neequaye et al., 2007), but these measures still neglect the contributing cognitive processes. Even presumably objective measures are difficult to define, however. For instance, defining and weighting errors for an error analysis is not a black-and-white issue (Neequaye et al., 2007). Evaluators may differ in the perceived seriousness or cause of an error.

Many of the task-oriented assessments described above share a focus on technical skills (Mitchell et al., 2014). Along with a focus on technical skill comes a tendency for evaluations to be very proceduralized and prescribed, fitting only one task rather than domain-level ability. The poor performance of many existing performance measurement tools may be due to a focus on readily apparent technical skills at the expense of nontechnical, more generalizable skills such as cognitive processes, decision making, and sociocultural behaviors (Bech et al., 2010). Focusing on the technical aspects of surgery and neglecting other aspects of surgical performance such as team processes can lead to assessment problems (Alderson, 2010). Criterion-based surgical performance measures focus on goal enactment, neglecting goal establishment. More process-based measures incorporating goal establishment processes will likely help improve surgical performance measures and allow them to make finer distinctions among surgeons to better capture experts along with trainees.

1.3.4 Issues with surgical performance measures.

Few of the tools utilized to measure performance in medicine demonstrate sufficient validity or correspondence with objective measures (Kogan, Holmboe, & Hauer, 2009). Operative log data generally fails to capture trainee understanding or quality of participation in a procedure (Mitchell et al., 2014). Written and oral exams only capture low-level knowledge (Mitchell et al., 2014). Out of the variety of assessments available that utilize direct observation of behavior, only two (the 7-item Global Rating Scale and the Procedure Based Assessment) received high grades from the ACGME medical education council (Jelovsek, Kow, & Diwadkar, 2013). As enumerated below, performance measures currently in use suffer from issues of discriminability, have difficulty capturing non-technical skills, are overly prescriptive, and leave opportunity for factors beyond surgical skill to impact scores.

1.3.4.1 Current measures do not effectively discriminate among various levels of skill.

Checklists are good for rating novices and can facilitate feedback, but they fare poorly at differentiating between experienced surgeons. Nonsurgeons are capable of evaluating surgeons using checklists, but cannot judge quality (Mitchell et al., 2014). Checklists are also vulnerable to problems arising from poor behavioral sampling, weighting/aggregation issues, and reliability/internal consistency among individual items. Global rating scales may yield more sensitive distinctions because expert raters are able to detect and incorporate subtle variation in performance and behaviors when allowed to give a general impression of performance rather than being confined by a checklist. Global rating scales allow the raters to incorporate quality into their ratings and are able to discriminate among all levels of performance, making them better for evaluating experts (Bech et al., 2010; Mitchell et al., 2014).

The most comprehensive solution may be to use a combination of the two approaches (Ahmed et al., 2011; Neequaye et al., 2007). Combination checklist-global rating scales are generally found to be both valid and reliable (Mitchell et al., 2014; van Hove et al., 2010), but are generally too procedure-specific or long to be practical (Mitchell et al., 2014). In addition, different authors have different ideas about what should be assessed for a given procedure (Mitchell et al., 2014). There is no systematic analysis or sampling of the domain due to inadequate theories of performance.

1.3.4.2 The surgical domain has difficulty capturing nontechnical skills.

Evaluating a holistic view of the surgeon is preferable as a means to capture a wider range of behaviors, but is difficult because measures of nontechnical aspects of performance are rare (Gelinas-Phaneuf & Del Maestro, 2013). Nontechnical skills may be just as important to performance as technical skills (Flach et al., in press). Surgeons have increasingly recognized the contribution of nontechnical skill to surgical performance, and nontechnical skill assessments are becoming more common (Mitchell et al., 2014). Unfortunately, the surgical domain has not incorporated such behaviors in a systematic way and may miss key components of performance or include irrelevant ones. A better understanding of the full range of constructs that contribute to surgical performance across a range of tasks and levels of expertise will facilitate definitions of criteria for ratings and help improve rating scales.

Nontechnical skills include such behaviors as communication, leadership, teamwork, briefing, planning, preparation, resource management, seeking advice and feedback, coping with stress, situation awareness, mental readiness, assessing risks, anticipating problems, decision making, adaptive strategy use, and workload management (Bech et al., 2010; Yule et al., 2006). Many of these behaviors such as planning, decision making, and adaptive strategy use encompass the goal establishment and goal enactment behaviors characteristic of expertise highlighted above as also missing from cognitive analyses. Others such as mental readiness or leadership seem vague and ill defined. Few studies have attempted to decompose these skills into their component behaviors, resulting in a lack of adequate rating systems for these skills (Yule et al., 2006).

In fact, despite the considerable variation in types of nontechnical skill behaviors, nontechnical skills are often considered as a single entity. This makes the contribution of any single skill difficult to assess (Hull et al., 2012). For instance, competence in neurosurgery is generally evaluated based on technical competence and "other skills", which is a catch-all term encompassing professionalism, communication, expertise, and collaboration (Gelinas-Phaneuf & Del Maestro, 2013). The terms under the "other skill" umbrella are difficult to define and weight within the same category, leaving raters to make their own judgments. Studies of surgical performance evaluation have neglected the influence of these judgments on performance scores, however.

1.3.4.3 Prescriptive evaluations limit measurement.

Many evaluations include checklists, which grade the trainee on whether they completed the necessary steps in a procedure but ignore the quality of the work (Mitchell et al., 2014). Not only does this approach exclude goal establishment and the context-adaptive goal enactment behaviors characteristic of expertise, it may actually penalize such behaviors by failing to recognize better performance and legitimate alterations to procedure. By predefining specific steps, errors, or explanations expected from trainees, checklists leave no room to account for trainees who use a different approach, commit different errors, or offer additional explanations. Even when rating guidelines are included, such as in the Likert-type scales, the criteria remain open to interpretation (e.g., the difference between “rarely” and “sometimes” may not be the same for every rater).

The prescriptive nature and scaling of the rating form lead to two effects. First, expert behaviors such as adaptive goal enactment and balancing constraints are excluded. The prescriptive format means that legitimate deviations from prescribed norms cannot be accounted for in ratings for individual items. Trainees cannot get credit for going beyond expectations and are difficult to penalize systematically for unanticipated errors not included in the checklist. Second, the cognitive burden on the raters is increased. The scaling problem and incomplete criteria force raters to incorporate these factors into scores using their own judgments and interpretations, increasing the likelihood of variance between evaluators. I believe that although the specific criteria do not leave room to account for adaptive behaviors, raters accommodate them in their global ratings. I examined this possibility with my analyses.

1.3.5 Factors beyond skill may impact scores.

In addition to the issues specific to medicine described above, the use of subjective performance measures in many domains leads to the possibility of scoring influences beyond the skill of the person being evaluated. Observations of performance in surgery are subjective (Mitchell et al., 2014). Raters are often untrained on the rating scale being used, which makes it difficult to use scales reliably and accurately (Kogan, Holmboe, & Hauer, 2009). Due to the lack of training and subjective nature of most assessments, raters must use their own idiosyncratic criteria when making performance determinations.

People form personal construct systems which are used to judge and anticipate events. These constructs may lead to individual differences in what raters perceive or look for when making performance judgments; valued behaviors may vary across raters (Borman, 1987; Wilson, 2010). Such idiosyncrasy does not necessarily mean that evaluators' global impressions are not useful, however. Army managers tend to show good agreement about the important elements of performance, even though different officers emphasized different combinations of these elements in their performance evaluations (Borman, 1987).

People judge performance based on the relative weights of multiple attributes. In the case of surgical performance evaluation, the attributes are either absent or poorly defined on the rating form. Evaluators are left to introduce their own attributes and incorporate them according to idiosyncratic belief systems. Issues of multi-attribute decision making, halo error, and prior experiences of the evaluators can all affect scoring.

1.3.5.1 Multi-attribute decision making may affect how evaluators form global scores.

Whether raters use checklists or global ratings, the resulting score constitutes an overall assessment across many types of behavior. This requires raters to utilize a form of multi-attribute decision making (MADM). MADM is a process by which people make judgments based on a number of attributes (Westenberg & Koele, 1994). Attributes are generally categorized as positive or negative, and may be independent, conflicting, or incommensurate with one another. Evaluators must reach a conclusion that balances between the positive and negative attributes based on their preferences (Stanujkik, Magdalinovic, & Jovanovic, 2013). Furthermore, raters must often evaluate attributes with imprecise information. This requires fuzzy reasoning, which deals with uncertain information (Manoharan, Muralidharan, & Deshmukh, 2011). Combining fuzzy principles with multi-attribute decision making allows evaluators to offer balanced, comprehensive, and accurate ratings.

Raters may use a variety of strategies within MADM to determine a rating. Decision strategies can generally be described as either compensatory or noncompensatory and as additive or nonadditive (Westenberg & Koele, 1994). Compensatory strategies allow low values on one attribute to be made up for by high values on another, while noncompensatory strategies do not. Additivity refers to whether attributes are combined via summation or some other method. Additive strategies tend to be compensatory while nonadditive strategies tend to be noncompensatory (Westenberg & Koele, 1994).

Evidence indicates that raters are able to change their decision strategy in response to changing task complexity, most often defined in terms of the number of attributes or number of alternatives to be considered (Timmermans, 1993). Generally, increasing task complexity leads to simplified (noncompensatory) decision processes or information search. Compensatory patterns apply particularly when the number of attributes is small or if the decision comprises a judgment rather than a selection (Timmermans, 1993). Depending on how the evaluators apply MADM strategies, global scores may be influenced by a variety of factors within the evaluation context such as the complexity of the rating form.

1.3.5.2 Halo error may affect scores.

The influence of rater cognition on scores is also illustrated by halo effects, where a rater's evaluation on specific performance subscales is influenced by some broader impression. Scores on a dimension of performance may reflect both actual observed performance and the rater's impressions of that performance (Solomonson & Lance, 1997). Halo is typically treated as error arising from the cognitive processes of the rater (Feeley, 2002; Murphy, Jako, & Anhalt, 1993). The result of halo error is to inflate correlations among different rating dimensions within raters relative to the true correlation (Dennis, 2007; Feeley, 2002; Murphy, Jako, & Anhalt, 1993). Halo can also lead to underestimates of the discrepancy across rating dimensions and underestimates of change in performance over time (Dennis, 2007).

Halo is generally considered to be composed of two components. The true halo component represents genuine overlap among rating dimensions, while the illusory halo component represents rater factors such as poor memory or poor measurement of behavior (Murphy, Jako, & Anhalt, 1993; Solomonson & Lance, 1997). Though originally defined as instances when a rater's general impression influenced specific judgments, halo has incorporated other definitions over time (Balzer & Sulsky, 1992). Halo error is now generally defined as one of three forms.

General impression halo occurs when a rater's overall impression of a target influences judgments of performance on independent criteria so that scores reflect both actual ratee performance and the rater's general impression (Fisicaro & Lance, 1990; Lance, LaPointe, & Fisicaro, 1994). The *salient dimension* model of halo error asserts that a rater's impression of performance on a salient dimension of performance may influence ratings on other dimensions (Fisicaro & Lance, 1990). This leads to error because raters' weights for criteria will vary and raters will have better opportunity to observe certain aspects of performance. Ratings will reflect actual target behavior as well as error based on the salient dimension (Lance, LaPointe, & Fisicaro, 1994). Halo due to *inadequate discrimination* occurs when raters fail to discriminate between dimensions of performance so that performance on one dimension influences scores on another (Fisicaro & Lance, 1990). Scores therefore reflect both target behavior and other (potentially irrelevant) dimensions (Lance, LaPointe, & Fisicaro, 1994). The general impression, salient dimension, and inadequate discrimination models of halo all have support in the literature (Dennis, 2007; Lance, LaPointe, & Fisicaro, 1994; Solomonson & Lance, 1997).

Several factors influence the occurrence of halo. Rater unfamiliarity with the target generally increases halo error (Feeley, 2002), but not always (Dennis, 2007). Halo tends to increase when the rater is not familiar with the person being rated because raters are more prone to using global impressions (Murphy, Jako, & Anhalt, 1993). Insufficient concreteness of rating dimensions either due to poor definition or poor rater training also increases halo error by increasing the need for rater interpretation (Feeley, 2002). Ratings of current or recent behaviors are less prone to halo because raters are not forced to rely on memory, which may be influenced by a global impression (Feeley, 2002; Murphy, Jako, & Anhalt, 1993). Poor rater motivation or effort can increase halo (Feeley, 2002). Using multiple raters for a single ratee can reduce halo effects, but at least five raters are usually required in order to see a benefit (Feeley, 2002).

Halo may result from poor rater motivation, poor observation, poorly designed rating instruments, or lack of rater training, all of which indicate areas for improvement in ratings (Jackson & Furnham, 2001). However, the presence of halo error does not necessarily diminish the value of rating scales (Murphy, Jako, & Anhalt, 1993). Global assessments may still be reasonably accurate even if halo is present (Jackson & Furnham, 2001). Evaluations are subject to influence from evaluator cognitive processes due to halo just as they are subject to influence due to MADM processes. Despite this, global rating scales can still provide insight into performance due to their overall accuracy even in the presence of halo effects.

1.3.5.3 Prior experience within and between candidates can affect scores.

In addition to individual factors such as leniency or criterion weighting during evaluation, raters can be influenced by contextual factors, particularly the structure of the rating form itself. Responses to items early in the rating form may influence responses to later items. For instance, halo effects may increase if early judgments about performance are related to global performance (Murphy, Jako, & Anhalt, 1993). Previously encountered items on a rating form provide context to later questions, shaping responses to those items (Tourangeau & Rasinski, 1988). Prior items help a rater determine how to interpret later items, as well as determine what is worth noting and what may be redundant. For example, if a survey question about general happiness follows a question about marital happiness, the respondent may exclude marital happiness from general happiness because that information has already been provided (Tourangeau & Rasinski, 1988). Prior questions may also serve as a benchmark standard or criteria, serving as anchors or points of comparison for future judgments (Tourangeau & Rasinski, 1988). The rating form may also influence responses via the activation of rater attitudes.

Attitudes may be retrieved, constructed based on context, or a combination of both (Tourangeau & Rasinski, 1988). Attitudes may be activated by the structure of an evaluation form. Attitudes can be thought of as long-term memory structures (Judd et al., 1991; Tourangeau & Rasinski, 1988). Attitudes can therefore form associations with one another (Judd et al., 1991). If one attitude is activated by a rating item, associated attitudes may have a greater chance of activation as well (Judd et al., 1991; Tourangeau & Rasinski, 1988). A series of studies by Judd and colleagues (1991) indicate that not only do previous items appear to activate attitudes for later items, but that expressing an attitude about a prior item may make responses to a related item more extreme. Context may serve to make future responses stronger in the priming item's direction or move responses in the opposite direction. The exact effect of context depends on the nature of the contextual items and whether respondents become aware of the context at a conscious level (Tourangeau & Rasinski, 1988).

Many variables can impact contextual effects within a rating form. The familiarity of the rater with the form and with the target of assessment influences how items are interpreted. The format of the questions, the complexity of the judgments required, and the interrelationships among items also affect responses (Tourangeau & Rasinski, 1988). Contextual effects are stronger when items are closer together (Tourangeau & Rasinski, 1988; Tourangeau, Singer, & Presser, 2003). Question order affects responses to questionnaires, but the correlation between items and overall validity of findings do not appear to be affected by context effects (Tourangeau, Singer, & Presser, 2003). Context effects are likely common, but confined to nearby conceptually related items within a questionnaire and do not impact the substantive findings or predictive validity of results (Tourangeau, Singer, & Presser, 2003).

1.4 The Current Study

The present study expands the current conceptualization of expertise to encompass a more comprehensive range of behavior and improve measurement techniques. Following the traditional paradigm of expert-novice comparison, I examine archival surgical audio/video data along with real-time ratings in journeyman and expert surgeons using both checklists and global ratings. Sampling both journeymen and experts provides a useful comparison that facilitates the identification of unique expert behavior and allows me to examine the developmental pathway of other behaviors. Though comparisons involving true novices may provide different insight into the data set, true novices with minimal background in medicine were not available here. I first identify higher-level constructs within these data to capture both currently-measured behaviors and additional to-be-included behaviors in performance metrics. I then examine the contributions of self-awareness and experience to such performance. Finally, I examine the influence of the raters on performance scores, in order to account for performance as completely as possible.

Performance ratings will allow us to explore the definition of expertise, specifically the relationship between experience and expertise. Video and think-aloud data will allow us to establish the importance of cognitive penetrability in skilled behavior and provide insight into how knowledge and action interact. These processes speak to larger issues within expertise theory, specifically the balance between cognitive control and automaticity in expert behavior. If the processes by which experts act in the world are cognitively penetrable, then expert behavior may not be as automatic as some researchers would suggest. More deliberative processes may be at work even in the presence of very rapid, seemingly intuitive decisions and behavior.

I will address how expertise is conceptualized and measured within both medicine and psychology, promoting a more process-oriented definition that includes the entire process of surgery from problem identification to execution. Though these findings are best characterized as descriptive, they serve to generate new hypotheses regarding the nature and measurement of expertise. By including concepts such as goal establishment and goal enactment, specifically self-monitoring behaviors, I will be able to examine the potential contribution of deliberate processes to various aspects of skilled performance. Whereas Robinson (2011) lacked standardized tasks or performance measures to link goal establishment and goal enactment to outcomes, this study allows me to establish such a link and demonstrate the importance of goal establishment and goal enactment to expert performance. I will also be able to comment on the nature of the interaction between goal establishment and goal enactment, further extending the work of Robinson (2011).

The remainder of the document is divided into chapters, each devoted to a specific aspect of my analyses (a complete list of my analyses can be found at the beginning of this document). Chapter 2 describes the methods and procedure of the parent study at the University of Maryland upon which my analyses are based. Chapter 3 describes the process by which I derived new variables to complement those already included in the parent study. Included in this chapter are analyses to check the interrater reliability of the captured variables, ensure sufficient variability in the performance data for analysis, validate subjective outcome measures against more objective indicators of performance, and a principal components analysis to group variables derived from audio and video data into conceptually related factors. In this chapter I also demonstrate the added explanatory value of a content analysis of the transcripts, as opposed to simple word count. Chapter 4 describes analyses related to establishing the importance of variables related to goal establishment and goal enactment in accounting for variance in global performance scores, particularly the role of analytical thought processes. I examine possible mediation between the existing variables of the surgical study and the new variables derived for this dissertation, as well as the generalizability of the newly identified variables compared to the existing surgical performance measures. Chapter 5 further establishes the contribution of conscious processes to performance via an analysis of the relationship between self-awareness measures and performance scores. I examine whether self confidence predicts performance, whether self confidence changes in response to past performance, and whether more experienced or more skilled surgeons are better able to judge their ability than less experienced or less skilled surgeons. Chapter 6 examines the link between experience and performance, challenging the idea that expertise is only developed over many years and indicating the utility of process-oriented performance measures. I examine whether a surgeons' years in practice predicts performance after accounting for training effects, and identify process-based differences between surgeons who on the surface share similar outcomes. Chapter 7 examines the influence of factors beyond the skill of the surgeon on performance scores, demonstrating that evaluator cognition impacts performance measures. I examine interrater reliability across levels of surgical skill and training status, as well as the effects of contextual features unrelated to the skill of the surgeon and the effects of demographic characteristics on scoring. Finally, in chapter 8 I summarize my findings and discuss the theoretical and methodological contributions of this work, along with limitations and directions for the future.

Chapter 2 – Method

This archival study relied on surgical performance data collected at the University of Maryland[1] as part of an evaluation of the Advanced Surgical Skills for Exposure in Trauma (ASSET) training course. Data included video recordings of procedures, subjective performance ratings, self-confidence ratings, and think-aloud protocols from trainees during procedures on human cadavers. In order to distinguish between the archival variables and the new variables introduced in the present study, I refer to the archival variables as “Maryland” variables and new variables as “WSU” variables.

The scope of the Maryland ASSET evaluation reflected the goals of the project and the logistical and resource challenges of recruiting experts in a specialized field. The present study served as a complementary analysis to better understand the underlying, general components of expert performance and the properties and vulnerabilities of the evaluation process. The Maryland data presented an opportunity to answer new research questions beyond the scope of the ASSET evaluation effort and contribute new understanding to the nature of expert conceptualization and performance, in the hope of better measuring expert performance in the future.

The ASSET evaluation was conducted under a grant to C. MacKenzie, with V. Shalin consulting on instrument design, data collection, and analysis issues. The ASSET

[1] All applicable IRB evaluations for the parent study were conducted under the auspices of the IRB at the University of Maryland, including the use of cadavers. For the purposes of this dissertation, the research was reviewed by the WSU IRB and determined to be exempt.

course consists of lecture-based instruction on multiple rarely-performed procedures in trauma management, with this specific evaluation focusing on four such procedures: exposure of the axillary artery, exposure of the brachial artery, exposure of the femoral artery, and lower leg fasciotomy. The present analyses focus only on the axillary artery exposure for multiple reasons. Of the four procedures, the axillary procedure offered a greater variety of strategies to attain the goal, providing greater variability in behavior to analyze. The procedure is also relatively short, making video and audio analysis and interpretation more tractable.

The remainder of this method section includes background information regarding the ASSET course and broader evaluation study in order to provide context for the analyses of the dissertation, followed by a description of each analysis. The measures, analysis process, and findings for each analysis will be described in turn.

2.1 ASSET Background

ASSET training is intended to improve and maintain Army surgeons' readiness to treat common battlefield casualties that rarely occur among civilians (limiting practice opportunities and leading to skill decay between deployments). The University of Maryland evaluation concerned the effectiveness of ASSET training immediately following the course. Evaluation focused on technical skills (e.g., how instruments are used), anatomical knowledge and identification, and the pace and efficiency of the procedure. The project also intended to improve the quality and efficiency of surgical skill evaluation by determining whether ratings of videotaped procedures are comparable to in-person evaluations. Additional goals included evaluating long-term skill retention, evaluating the effectiveness of training using a surgical model compared to a cadaver, and developing a software tool to assist in assessing surgical performance, but these aspects of the evaluation were not addressed directly in the dissertation.

2.2 Participants

2.2.1 Observed surgeons.

A total of fifty practicing surgeons performed procedures as part of the ASSET evaluation. The less experienced subset included residents and fellows. Though these surgeons were not true novices (they had medical training and some level of supervised experience), they were not licensed to practice independently. Residents included 24 men and 12 women with between two and five years of experience. Fellows included 1 man and 3 women, all with six years of experience. Two of the residents and two of the fellows were left-handed.

An additional 10 attending surgeons were recruited due to skilled reputation. These surgeons were licensed to practice independently. The attending surgeons included seven men and three women with between two and 33 years of experience post-residency. One attending surgeon was left-handed.

Despite the fact that the participants in this study were all qualified surgeons, even the surgeons with the most experience overall may have had little experience with the axillary artery exposure. Surgeons in residency only perform an average of around two major vascular repairs for trauma of any type. These procedures are equally rare in daily practice – the reason these procedures were studied during the ASSET evaluation is that Army surgeons rarely perform them between deployments and need to maintain their skills.

Of further note is the sample size for this study. Though some of my later analyses may suffer from limited power or sampling issues (discussed where relevant, and also in Chapter 8), the parent Maryland study represents one of the largest controlled, whole-task, open surgical studies available. Most studies with comparable sample sizes have used part-task or laparoscopic stimuli (Mackenzie, personal communication). This study therefore provided an excellent sample for the surgical domain.

2.2.2 Evaluators.

A total of 18 evaluators provided performance evaluations for the ASSET study. Evaluators were experienced surgeons or specialists with advanced training in a related field such as anatomy. The evaluators included 10 men and eight women, with between two and 47 years of experience in their field of expertise. Due to uncontrollable availability, the range of evaluations performed by each evaluator was from one to 42. Only six evaluators performed more than 10 evaluations, and only three evaluators performed more than 20 evaluations.

2.3 Apparatus and Surgical Cadavers

Participants wore a head-mounted camera to record the trainees' actions and audio during the procedures. An additional camera was mounted above the surgical site to provide an overhead view and capture additional audio. Participants had access to standard surgical tools during the procedure.

Participants performed the surgical procedures on recently deceased, unpreserved cadavers. The cadavers were a mix of generally elderly males and females with a range of body types. Both left and right sides of the cadavers were used.

Participants also performed procedures on surgical models designed to simulate the human body. These models were used as part of a concurrent effort to evaluate the effectiveness of these models for surgical training.

2.4 Maryland Study Design

The Maryland parent study followed a pre-test, post-test training design examining the impact of several factors on surgical evaluation. For the purposes of this dissertation, the pre-test procedures, post-test procedures, and attending surgeons' procedures served to maximize the observed range of performance. I describe this design here to facilitate later discussion of performance measures and data analysis. The 40 less-experienced surgeons were tested before and after ASSET training in order to assess the immediate effectiveness of the course, generating 80 training procedures in total. The 10 attending surgeons were evaluated for a single procedure, bringing the total number of available evaluations used in my analyses to 90. All aspects of the study design involving the residents were within-subjects. All comparisons involving the attending surgeons were between-subjects.

As mentioned, the surgeons performed four separate procedures. Trainees performed pre-ASSET evaluations for all four procedures together, received ASSET training, and then performed post-ASSET evaluations for all four procedures. The data procedures (attending surgeons, pre-training residents, and post-training residents) were evaluated sequentially with the attending surgeons evaluated first, then the pre-ASSET training procedures, and finally the post-ASSET training procedures.

Post-ASSET evaluations involved a surgical model in addition to the cadaver. Surgeons performed all four procedures on both the model and cadaver. All eight procedures were performed on the same day, the order of which was randomized within the model and cadaver. The same evaluators judged both the cadaver- and model-based post testing procedures. Trainees answered the clinical portion (non-surgical performance elements addressing suspected injuries, patient examination, additional tests, resuscitation plan, etc.) for the first set of procedures only. For example, if the surgeon performed the first post-ASSET procedure on the surgical model, the full evaluation would be used. For the second post-ASSET (cadaver-based) procedure, the surgeon would immediately start conducting the physical procedure on the cadaver (beginning with identifying landmarks and marking the incision) and forego the earlier portions of the evaluation. Because I wanted to ensure comparability with the pre-training procedures, I only analyzed data from the cadaver-based procedures. This design element therefore necessitated that I sometimes had to combine clinical questions from a model-based procedure with physical action from the cadaver-based procedure in order to obtain a full data set for some post-ASSET procedures.

2.5 Maryland Performance Evaluation

Two raters typically accompanied trainees during the hands-on procedure for both pre and post testing, although in some cases only a single rater was used. One rater was usually an experienced surgeon and one was a non-surgeon with medically related experience (e.g., graduate training in human anatomy). The raters each completed separately a single real-time evaluation for the procedure based on the trainee's clinical performance (impressions of the nature of the patient's injury, plan to diagnose the injury, and care plan), and surgical performance. Due to the scheduling of co-located, simultaneous training activities, these raters were necessarily aware of trainee experience and whether the procedure occurred before or after training. Early ratings employed a paper form while later ratings employed an electronic tablet-based form.

All procedure performance ratings in the parent Maryland study were generated using the same rating criteria. The sheet consisted of a combination of checklist-style yes/no evaluations for whether the trainee listed certain concerns, ordered certain images, made any errors, and completed different steps of the procedure. Both clinical (i.e., focused on diagnosis, patient assessment, and care plan) and surgical (i.e., focused on the technical execution of the procedure) aspects of care were addressed in the form. The clinical segment of the form was not evaluated as part of this dissertation because many of the available recordings of the procedures omitted this portion of the evaluation and thus we could not transcribe it for our own analyses.

The surgical portion of the form utilized a total of 35 items. Twenty items were yes/no items related to completing steps in the procedure, elements of operative technique, and elements of instrument use. Ten items evaluated technique points on a 5-point Likert-type scale. Overall ratings of clinical skill, anatomical knowledge, surgical technique, and readiness to perform the procedure were provided using 5-point Likert-type scales. Finally, the evaluators provided a global rating score from 0-100 based on the same anchors as the overall readiness Likert-type scale. The script and rating form used by the in-person raters is provided in Appendix A.

The Maryland study utilized an additional outcome measure calculated based on a summation of subcomponents of the rating scale. This measure excluded errors and completion time primarily due to scaling considerations. Called the Individual Procedure Score (IPS), it represented a ratio of observed performance to maximum performance. IPS scores were intended to serve as an overall measure of performance, similar to the 0-100 global score. Scores in each area of assessment (overall knowledge, anatomic knowledge, patient management knowledge, procedure steps, and technique) were calculated as a percentage of total available points. The IPS was calculated as the sum of the total points earned, divided by the sum of possible points.

2.6 Procedure

As mentioned, the dissertation focused on the axillary artery procedure, but the parent training study investigated four separate procedures. Pretesting and post testing for each procedure followed the same basic structure, with procedure order determined by a Latin square design. The axillary artery exposure thus occurred in the context of other procedures, but stands alone as its own operation. The protocol for evaluating axillary artery management is described below.

2.6.1 Pretesting.

ASSET trainees completed a pretest prior to receiving ASSET training. Pretesting consisted of self-ratings of confidence as well as a hands-on evaluation. Participants first rated their confidence in their ability to perform individual aspects of the procedure. They then performed an axillary artery exposure as one of four cadaver surgeries while being evaluated.

The hypothetical case was presented as a 24-year-old male with a gunshot wound to the chest (the wound was on the patient's left side for some procedures, and on the right side for others). Prior to the procedure, trainees received the case history of the hypothetical patient and were asked to diagnose possible injuries. Trainees were then asked what physical findings they would look for to determine the nature of the patient's injuries, along with any additional imaging or studies they would seek. The results of the physical exam and imaging were presented to the trainees, and the trainees reported their plan for the patient, including initial resuscitation[2]. Participants then demonstrated how they would position the patient on the operating table to best perform the procedure.

[2] Participant responses may have been biased by prior expectations, as the design of the study did not allow for real alternative diagnoses or superfluous test results. Prior stages of the evaluation therefore likely guided the care plan.

Following this explanation, the trainees used a marker to identify anatomical landmarks on the patient and the incision they planned to use for the procedure. Finally, the trainees performed the procedure with the goal of gaining control of the axillary artery with a vessel loop on the proximal (towards the center of the body) side of the bullet wound to stop hypothetical bleeding. Trainees were allowed a maximum of 20 minutes from their first cut to complete the procedure, under the rationale that the patient would have exsanguinated (bled to death) after 20 minutes.

Trainees started the procedure standing on the same side that was “wounded”. The surgeons were free to reposition themselves as needed during the procedure and could adjust lighting or the operating table as necessary. ASSET evaluators provided assistance in the form of handing trainees instruments or providing a second set of hands when needed (generally serving to hold things in place). Participants were instructed to think aloud during the skin marking and procedure. In the event that the trainee stopped talking the evaluators prompted them to speak aloud. Though not part of the official instructions, a commonly used directive was to ask the trainee to envision that the evaluators were first year medical students and the trainee was explaining the procedure to them. The team of two evaluators filled out the evaluation form in real-time during the procedure, as described previously.

Immediately after the procedure, evaluators asked participants about the consequences of ligating the axillary artery and some of the common pitfalls during the procedure. Participants also received brief feedback related to whether the procedure was performed well or how the trainee could have done better. Raters minimized instruction during feedback; feedback for the axillary procedure generally focused on the approach taken to reach the artery. Trainees then provided a post-procedure confidence rating.

2.6.2 Training.

After pretesting, trainees received the ASSET training course. The ASSET course consists of instructor-led, lecture-based training along with hands-on cadaver work. Lectures are conducted in a group setting, typically with 20 to 40 trainees per class. Up to four trainees share a cadaver. The course lasts a single day and covers 47 procedures during that time, including the four assessed for the parent study. Training for the axillary artery procedure in question covered approximately six pages of the 154-page manual and required approximately 10 minutes. Course content for the axillary procedure included bullet points offering general guidance for preparing the patient, broad descriptions of anatomy, and common pitfalls to avoid.

Due to the uncontrollable intervening professional activities of the trainees, some trainees may have received additional hands-on experience with a procedure prior to post-testing. Although I did not have a way to assess this possibility, the uncommon nature of such procedures in normal surgical settings minimized the risk that additional experience affected the results of the study.

2.6.3 Post testing.

Post-ASSET testing occurred four weeks after the training sessions. The posttest procedure was the same as the pretest, including the same case presentation. As part of the parent study comparing surgical models to cadavers, trainees performed the procedure twice during post testing – once on a surgical model and once on a cadaver.

Chapter 3 – Variable Identification

Data available for my analyses included:

1. Demographics (e.g., experience, prior training/courses, etc.)
2. Four sets of trainee self-ratings of confidence:
    a. Before and after the pre-ASSET axillary procedure
    b. Before and after the post-ASSET axillary procedure
3. Pre-training global and subscale performance ratings
4. Post-training global and subscale performance ratings
5. IPS scores
6. WSU predictors identified from video and audio of the procedures

Items 1 – 5 result from the parent Maryland study. The Maryland performance measures (3 – 5) relied upon multiple evaluators, while the WSU measures introduced in this study relied on only a single rater. As a result, I needed to condense the Maryland ratings into a single score in order to provide a one-to-one ratio with the WSU performance measures. I elected to achieve this by combining the performance ratings across evaluators for the Maryland measures. In the case of binary items, cases of rater agreement that an action occurred were scored a "1", cases of disagreement were scored a "0", and cases of agreement that an action did not occur were scored as "-1". Likert-type, interval, and continuously valued items were averaged between the two raters. When scores from only one rater were available, that rater's values were used.

3.1 Coding the Data

Qualitative data from the procedure videos and trainee think-aloud protocols served to identify WSU behaviors to explain performance ratings, as well as identify new constructs to capture unexplained variance in global scores. The time investment required to train new raters, coupled with high turnover rate among undergraduate research assistants, made training secondary coders impractical. I therefore elected to code all of the data myself. Videotapes of the procedures were coded using the head-mounted camera as the primary source of data, with the overhead camera used as a backup in the event of poor audio or an occluded visual field. Transcriptions of the think-aloud protocols were similarly coded. I then examined reliability among the Maryland evaluation items and the enumerated WSU variables, eliminating those that were not reliable. I performed data reduction on the WSU variables using a Principal Components Analysis (PCA) to derive new constructs, and validated the Maryland global score that served as the primary outcome measure for my analyses.

3.1.1 Developing the coding scheme.

The procedure videos and think-aloud protocols each had separate coding schemes. I developed the WSU coding schemes in the spirit of grounded theory using an iterative process guided by the constructs of goal establishment and goal enactment. I identified candidate behaviors a priori based on the goal establishment and goal enactment behaviors identified by Robinson (2011). In addition, I watched videos of several surgical procedures and listened to think-aloud protocols from the trainees and think-aloud recordings of evaluators watching videos of the procedures. I watched four axillary artery exposures, two brachial artery exposures, and two fasciotomy videos, as well as listened to the trainee think-aloud protocols from the same procedures.

Preliminary evaluation of the videos and think-aloud protocols allowed me to become familiar with the domain and identify behaviors that appeared important but were not listed in the formal evaluation form, as well as behaviors that appeared important to raters during their evaluations. I also identified behaviors that appeared to be related to the subjective items on the evaluation form in order to explore rater cognition and facilitate operationalizing subjective rating items.

After generating a preliminary set of candidate behaviors, I applied the WSU coding scheme to the videos and think-aloud transcripts. As I coded, I modified the scheme by adding or altering behaviors to capture greater detail. Behaviors were removed that did not appear to show variance between trainees or evaluators, or that could not be reliably defined and identified. This followed an iterative process until a stable coding scheme emerged that appeared to capture relevant behaviors and could be applied consistently.

3.1.2 WSU coding scheme overview.

All codes were time stamped to facilitate matching across the video and think-aloud data. In order to synchronize across the two sources of data, time stamps for both the audio and video codes were based on elapsed time (in seconds) since the beginning of the procedure (identified as the beginning of the instruction to describe and mark on the skin the landmarks and incision the surgeon planned to use). Audio transcripts were stamped based on the beginning of each speaker's utterances, the beginning of a new idea from the same speaker, or as needed to help maintain a sense of elapsed time for particularly long utterances. Videos of the procedures were coded by the second: I constructed a spreadsheet with the total number of seconds for a given video in the first column and the individual behaviors in each subsequent column. Behaviors occurring at any given second within the video were marked in the respective column.

In addition, each procedure was subdivided into four phases: incision, muscle, identification, and control. The *incision* phase of the procedure was defined as the time from the knife first contacting the skin of the cadaver to the time that the trainee had reached the muscle tissue beneath. The *muscle* phase of the procedure was defined as the time from reaching the muscle to successfully dividing the muscle and reaching the vascular structures underneath. The *identification* phase of the procedure was defined as the time from reaching the vascular structures to identifying the axillary artery. Finally, the *control* phase of the procedure was defined as the time from identifying the artery to clamping the vessel loop to obtain control.

These phases generally occurred in a linear order, but trainees could move back and forth between them in the case of extending the incision or if the artery was misidentified and the trainee had to continue searching. In addition, although these phases are objectively defined, they were identified partially based on the trainee's perception of events rather than actual events because trainee perception drove behavior. For instance, if the trainee thought they had identified the artery but had actually identified a vein, the control phase would still begin at the point at which the trainee started working on the vein.

All individual behaviors from the procedure videos and think-aloud protocols were placed into subcategories within goal establishment and goal enactment in order to ensure that the coding scheme could capture constructs of interest and to facilitate implementing the coding scheme. This categorization was based on hypothesized rather than empirical results. Final categorizations based on the data analysis are described in the results section. The initial categorization described here merely served to help structure the coding scheme.

Three broad categories served as the guiding structure for all other categories: goal establishment, goal enactment, and rater cognition. Goal establishment behaviors were largely verbal behaviors derived from the think-aloud protocols, while goal enactment behaviors were both verbal and nonverbal derived from the videos. Each of these main categories in turn had several subcategories that served to capture the individual observed behaviors:

3.1.2.1 Goal establishment.

- *Problem detection* behaviors indicated that the trainee had made a mistake or gotten lost. The trainee may or may not have recognized the error.
- *Problem anticipation* behaviors indicated that the trainee had anticipated a possible issue before it occurred.
- *Planning* behaviors were related to deciding how best to proceed once a problem had been diagnosed.
- *Monitoring* behaviors indicated that the trainee was either monitoring themselves or their progress through the procedure.

3.1.2.2 Goal enactment.

- *Adaptation* behaviors indicated that the trainee performed an idiosyncratic behavior or had to adapt their preferred method of operation to account for the immediate context. These were verbally indicated in the think-aloud protocols and observed in the videos.

- *Environmental control* behaviors helped the trainee alter the environment to best facilitate success. These were observed in the videos.
- *Technical aspects* were related to the technique points of surgery, such as instrument selection or dissection technique. These were observed in the videos.
- *Navigation* behaviors were related to moving through the cadaver and remaining oriented to find the target vessel. These were verbally indicated in the think-aloud protocols and observed in the videos.
- *Balancing constraints* helped the trainee prioritize competing goals such as speed vs. accuracy. These were verbally indicated in the think-aloud protocols.

3.1.2.3 Rater cognition.

- The *MADM* category identified contextual factors that were not directly related to surgical skill, but may have influenced scoring. These were derived from the think-aloud protocols and observed in the videos.
- *Halo* behaviors were behaviors that may have influenced an evaluator's overall impression of the trainee but weren't directly related to surgical performance or the scoring criteria. These were derived from the think-aloud protocols and observed in the videos.

3.2 Reliability Check

After coding the data set, I arrived at the final set of Maryland and WSU variables by eliminating unreliable measures. I selected Krippendorff's alpha as the best measure of reliability for this study because it is suitable across multiple measurement scales (i.e., categorical, ordinal, interval, ratio), can be used for any number of coders, and can accommodate missing data (Artstein & Poesio, 2008; Hayes & Krippendorff, 2007; Krippendorff, 2004). These qualities made Krippendorff's alpha best suited for allowing me to make direct comparisons across the variables in this study with a single reliability metric. The different evaluators used during the original University of Maryland study were compared within each procedure across the entire sample of procedures. As I was the only coder for the video- and transcript-based WSU measures in this study, I assessed reliability for these measures by recoding a subsample of the procedures. I recoded 10% of the sample procedures (nine procedures: four pre-ASSET procedures, four post-ASSET procedures, and one expert procedure), with a minimum of one month between coding sessions for a given procedure. I then computed Krippendorff's alpha using the procedures within this subsample.

An alpha of 0.80 is generally considered to be the benchmark for good reliability, but values of 0.60 are also acceptable in some cases (DeSwert, 2012). I chose to accept 0.60 as the cutoff for reliable measures as this was an exploratory study focused on hypothesis generation and I wanted to ensure that I was able to draw upon as many predictors as possible while still rejecting clearly unreliable variables. Further, Krippendorff's alpha can be low despite few instances of disagreement in the case of rare variables, particularly with small samples or binary measurement scales (DeSwert, 2012). Because many of our variables were binary and/or relatively uncommon and I used a relatively small recoding sample for the WSU video and transcript variables, I felt a lower alpha threshold would be better suited to the data set[3]. Krippendorff's alphas for all WSU variables are listed in Appendix B; Maryland variables are listed in Appendix C. All values were computed using the SPSS macro described in Krippendorff (2011).

[3] Variables determined to be reliable or unreliable here may differ from those in other published research utilizing the same data due to different reliability measures and my adjustment of the reliability criteria to accommodate the unique nature of the WSU variables.

Variables that did not meet the threshold of 0.60 were combined with related variables to improve reliability or were dropped from further analysis. Among the Maryland data, 60 out of the 106 individual evaluation items proved to be unreliable. Among the WSU variables, 12 out of 71 variables proved to be unreliable. As many of the unreliable Maryland items were clustered within the predefined sections of the Maryland rating form, I elected to combine the subitems within each of these sections in an attempt to improve reliability. Variables that remained unreliable were dropped, leaving only the reliable variables for analysis. The unreliable WSU and Maryland variables are discussed in further detail in Appendix D.

3.3 Examining Variability

I examined the variance of the remaining variables in order to ensure enough variability in scores to predict performance. I examined histograms of all variables in order to assess the distribution and variability of scores on each variable. Variables with the same score for greater than 90% of the surgeons were excluded from further analysis. All of the variables met our inclusion threshold for variability in the data set.

3.4 The Final Variables

The final set of variables used in all subsequently described analyses is described below. The Maryland and WSU variables are described separately, within the goal establishment and goal enactment framework described above. The final empirical grouping of the variables is described in the results section.

3.4.1 Maryland variables.

Below I describe how the Maryland variables group into goal establishment and goal enactment. Though at first glance both goal establishment and goal enactment appear to be included in the Maryland performance evaluation, the goal establishment variables were only related to identifying a diagnosis or recognizing problems ahead of time. Maryland variables related to the procedure itself were grouped into goal enactment – consideration of goal establishment is excluded from skilled task execution.

3.4.1.1 Goal establishment.

- *Question 1: Suspected injury* (section 1). The eight items in this section assessed specific anatomical structures that the surgeon suspected could have been injured. Each item in this section was scored on a binary yes-no rating scale, giving a maximum score of eight on this aggregated measure.
- *Question 3: Additional studies* (section 3). This variable is composed of the six items in section three of the Maryland evaluation form. These six items assessed the imaging that the surgeon would use to help reach a diagnosis, again on a yes-no scale. The maximum score for this aggregated measure was therefore six.
- *Question 12: Pitfalls* (section 12). This variable is composed of the five items in Section 12 of the Maryland evaluation form. The items in this section evaluated the surgeons' knowledge of common mistakes or problems that might be encountered during this particular procedure, again using a binary scale. The maximum score for this aggregated measure was five.

3.4.1.2 Goal enactment.

- *Question 7: Landmarks and incision* (section 7). The four items in this section assessed the anatomical landmarks that the surgeon would use to guide them, as well as the incision that the surgeon would make for the procedure. These items used a binary scale, leading to a maximum score of four for this measure.
- *Question 8, Part 1: Steps of the procedure* (section 8, part 1). This variable is composed of the seven items in section 8, part 1 of the Maryland evaluation form.

The items in this section evaluated whether the surgeon completed the proper steps of the procedure, again using a binary yes-no scale. The maximum score on this aggregated measure was seven.

- *Question 8, Part 2: Technique* (section 8, part 2). This variable is composed of the 10 items in section 8, part 2 of the Maryland evaluation form. The items in this section evaluated different elements of good operative technique using a 1-5 Likert scale, leading to a maximum possible score of 50 on this aggregated measure.
- *Question 9: Expert operative field maneuvers* (section 9). This variable is composed of the six items in Section 9 of the Maryland evaluation form. The items in this section evaluated elements of operative technique thought to distinguish expert surgeons from novices. Each item in this section was scored on a binary scale, giving a maximum score of six for this aggregated measure.

3.4.1.3 Additional measures.

- *Demographics.* Participants' age, sex, career status (resident, attenting, or fellow), and years of experience were captured as part of the Maryland study. Similar information was captured for evaluators as well.
- *External training.* The number of hours spent in the cadaver lab, open skills lab, and minimally invasive skills lab since medical school both before and after ASSET training, as well as whether the participant had taken any cadaver-based courses since medical school.
- *Confidence ratings.* Participants provided self-ratings of their confidence in their anatomical knowledge of the shoulder/axillary region, arm, forearm, inguinal

region, and lower extremity. They also rated their confidence in their ability to complete procedures in each of these regions. Confidence ratings were obtained before and after each procedure (pre and post ASSET), for a total of four sets of ratings. Ratings utilized a 1-5 Likert-type scale.

- *Cadaver body habitus.* This variable describes the body type of the cadaver (e.g., thin, average, or obese). Because weight is a continuum and because there was no good way to average the raters' evaluations, evaluators' ratings were combined into an ordinal variable such that if both evaluators agreed that the cadaver was thin, the variable was scored as a 1. If one evaluator said the cadaver was thin and the other rated the cadaver as average, the variable was scored as a 2. If both evaluators rated the cadaver as average, the variable was scored as a 3. If one evaluator said the cadaver was average and the other rated the cadaver as obese, the variable was scored as a 4. Finally, if both evaluators rated the cadaver as obese, the variable was scored as a 5.
- *Overall understanding of anatomy.* This variable represented a 1-5 global assessment of the surgeon's understanding of the anatomy of the Axillary region.
- *Overall readiness.* This variable represented a 1-5 global assessment of the surgeon's overall readiness to perform an Axillary artery exposure.
- *Global score.* This variable represented a 0-100 global assessment of the surgeon's overall performance. This variable served as our primary outcome measure in subsequent analyses.
- *IPS (Individual Procedure Score).* This variable is calculated from a selection of items in the Maryland evaluation form that represents the proportion of possible

points earned for those items. It is intended as an overall metric of performance, similar to the global score.

- *TRI (Trauma Readiness Index)*. This variable is calculated similarly to the IPS score, but accounted for performance across all four procedures assessed during the Maryland ASSET study. The TRI score is intended as a metric of general surgical ability rather than an assessment of skill on any particular procedure.
- *Critical technical error.* This variable was a binary assessment of whether the surgeon committed any technical error that would have killed the patient. These errors included failing to control the artery by misidentifying the structure, or by failing to finish within the time limit.

3.4.2 WSU variables.

I derived additional variables from the video and think-aloud data generated during the procedures, encompassing aspects of performance evaluation relevant for this study: goal establishment, goal enactment, and evaluator cognition. Variables from the think-aloud protocols are particularly interesting, as the ability to verbalize intent and action speaks to the cognitive penetrability of such behavior and the ability of surgeons to monitor and intervene in their own performance. The reliable variables are described here, organized here based upon my hypotheses of whether they belong to goal establishment, goal enactment, or evaluator cognition, and how they are related to each category. Later PCA helped confirm these groupings. Each variable was scored as a total tally during the procedure, unless otherwise noted. Some variables were also scored within phases of the procedure, and were only reliable within certain phases. These instances are described for each variable where relevant.

Although in some cases the full sets of WSU and Maryland variables captured similar actions (e.g., instrument use or instrument changes), many Maryland variables proved to be unreliable and were therefore not included in the final set of analyses. Similarly, many of the WSU variables were either unreliable or dropped out during principal components analysis (discussed later). These WSU variables likewise were not included in the final analyses (see Appendices 2 and 3). This dropout eliminated much of the overlap between the two sets of predictors. Any remaining overlap will be addressed in the results in section 4.2.1.3.

3.4.2.1 Goal establishment.

Problem detection

- *Realizing a mistake*. The surgeon noted some type of error, such as a navigation error, injuring a structure, or misidentifying a structure. For example, "I totally destroyed the enominant vein on this side."
- *Altering a vessel loop.* This action occurred after the trainee had already identified and controlled what they believed to be the artery, but later acted to adjust the loop (e.g., by removing it, loosening it, or moving it to another area on the vessel).

Planning

- *Forms a plan.* The trainee verbally identified intended action. For example, "I'm gonna leave this loop on but loose. I'm gonna put it on another right angle."
- *Weighing options.* The trainee was verbally comparing or deciding between multiple possible courses of action. For example, "Do I wanna go above or below the clavicle? I think I wanna be above the clavicle."

Monitoring

- *Expressions of doubt or uncertainty.* The trainee declared an absence of confidence in the ability to perform the procedure. For example, "This is pec major, and pec minor somewhere too. I'm not sure which is which."
- *Expressions of confidence or certainty.* The trainee declared positive affect regarding current experience. For example, "I see some vasovasorum, which makes me feel good about it."
- *Checking by feel.* The trainee touched or interacted with the body to gauge the status of the procedure and determine progress (e.g., determine whether they had completely divided a muscle).
- *Mentioning things they expect to happen.* The trainee declared expectations or anticipation. For example, "I'm looking at the vein over here and the artery's gonna be just behind it."
- *Double checking behaviors.* Trainees confirmed a vessel was the artery after looping it (e.g., by identifying other nearby structures). For example, "There (finished with the procedure). But let's dissect it out to be sure."
- *Evaluating progress.* Verbal indications of keeping track of the status of the procedure and how well the trainee was moving towards the goal. For example, "Getting through the parietal pleura bluntly with my right angle. Eh, I'm not, I'm not quite there yet."

3.4.2.2 Goal enactment.

Adaptation

- *Miscellaneous oddities.* Actions that appeared relatively unique to the trainee. Examples include making an incision through the armpit instead of the chest, or making an extremely large incision. This variable has some overlap with the Maryland items within Q8S1 and Q9, mostly due to the potential influence of the size of the incision. However, the Maryland variables are broader and also incorporate other factors including how well the surgeon utilized the available incision space and how efficiently the surgeon moved through the steps of the procedure.
- *Accounting for individual anatomy.* The trainee mentioned something unique about an individual patient such as being particularly thin or having scar tissue from a prior procedure.
- *Workarounds.* Workaround statements were related to mentioning things that the trainee would normally do or prefer to do but couldn't because of the constraints of the testing task. The most common example was the stated desire to use an electrocautery knife (a means of minimizing bleeding when cutting; not provided to the trainees during cadaver procedures) rather than a scalpel.

Environmental control

- *Environment adjustment (patient).* The trainee adjusted the positioning of the patient in order to facilitate the procedure.
- *Environment adjustment (workspace).* The trainee made the environment easier to work in, such as adjusting the operating table or repositioning a light.
- *Environment adjustment (self).* The surgeons repositioning themselves in order to better access a structure or work more comfortably.

- *Placing retractors or holding the incision open*. The trainee's behavior helped maintain an open work area in the patient and improve visibility.
- *Repositioning retractors*. The trainee altered retractors that had already been placed in order to improve visibility further or facilitate work in a new area.
- *Laying out instruments ahead of time*. The trainees selected a handful of instruments and organized them prior to beginning the procedure, presumably in order to have easier access later.
- *Extending the incision*. Some trainees had to make their incision larger to continue working. This usually occurred when the original incision was too small or was not located in the right place.
- *Risk mitigation.* The trainee acted in order to ensure smooth execution of the procedure. For example, the trainee may have explained the use of a certain technique or tool in order to reduce the risk of inadvertent damage to a vessel, such as "I'm hoping that by staying right on the clavicle I can stay away from important nerves."

Technical aspects

- *Time between identifying the artery and placing the vessel loop.* The number of seconds between identifying the artery and clamping the vessel loop.
- *Time spent searching for instruments.* The total number of seconds the trainee spent looking for instruments during the procedure.
- *Instrument changes.* This behavior was indicated when the trainee changed instruments or picked up a new instrument for the first time. This action was scored both as a tally during the procedure as a whole and during individual

phases of the procedure. This variable was reliable for the individual phases of the procedure as well as for the procedure as a whole.

- *Proportion of the time using instruments in two hands.* The number of seconds the trainee worked using instruments in both hands during a particular phase of the procedure, divided by the total amount of time required for that phase. This measure was scored within each phase of the procedure, and proved reliable for all four phases of the procedure.
- *Shifts between blunt and sharp dissection.* The number of times the trainee changed dissection strategy between blunt and sharp dissection. Such a change may or may not have been associated with an instrument change, as some instruments can be used for both strategies.
- *Proportion of the time using blunt and sharp dissection.* The proportion of the total amount of active dissection time that blunt and sharp dissection were each used. This behavior was scored within each phase of the procedure. The proportion of the time that the surgeons used sharp dissection was reliable for all four phases of the procedure, while the proportion of the time the surgeons used blunt dissection was only reliable for the muscle and identification phases of the procedure.
- *Completion time.* This represents the number of seconds required for each phase of the procedure, as well as the total completion time. All phases of the procedure were reliable except for the incision phase. Total completion time was also reliable.

- *Idle time.* The percentage of the total procedure time that the trainee was not engaged with the patient or otherwise occupied (such as looking for an instrument). Idle time was largely hesitation where the surgeon paused to think or removed an instrument from a structure and was slow to transition to another structure.
- *Backtracking.* The number of times the participant had to revisit steps of the procedure (e.g., extend an incision or go back to identification after entering the control phase).

Navigation

- *Naming structures.* Verbally identifying structures by name either to remain oriented or as part of forming a plan.
- *Knowledge.* Verbal indications of navigation using specific knowledge of anatomy based on technical criteria or other definitions. For example, the axillary artery is actually a section of one longer blood vessel that extends all the way into the arm. The section considered to be axillary artery is marked by anatomical landmarks: "Once it hits pec it changes over to subclavian so axillary - this is technically axillary artery."
- *Exploration.* Interacting with the body to try to find something familiar or gain a sense of where the trainee was working. This behavior was scored as total time (in seconds) during the procedure.
- *Evaluating structures (by feel).* Identifying a structure by touching it to see if it was tubular, etc.

- *Evaluating structures (observable bodily behavior).* The trainee mentioned the ability to identify structures based on pulsation, etc. For example, “The artery’s gonna be right here and it would be pulsing in real life.”
- *Evaluating structures (location).* Verbal identification of structures based on where they are in the body. For example, “I’m not entirely sure that’s not the carotid…it kind of looks like it’s going up into the neck.”
- *Heuristics.* Verbal declarations of navigating using general knowledge of the body (e.g., arteries tend to be located deeper in the body than veins). For example, one is likely to encounter the vein prior to the artery in the body: “I’m looking at the vein right here and the artery’s gonna be just behind it.”

Balancing constraints

- *Balancing constraints.* The trainee discussed competing goals or prioritizing actions during the procedure. For example, “Um (sigh) I don't really wanna take these large vessels. If this guy was bleeding incredibly I would just take this shit. I might take a smaller trail branch.”

3.4.2.3 Rater cognition.

MADM

- *Surgeon’s dominant hand.* The surgeon’s preferred hand, identified by which hand was used to hold the marker and scalpel.
- *The side of the cadaver on which the procedure occurred.* This was coded based on whether the surgeon stood on the cadaver’s left or right side. Combined with the surgeon’s dominant hand, this allowed us to examine whether different combinations of handedness and operating location made the procedure easier or

harder (due to the reaching motions necessary) and may have affected scores depending on how the evaluators took that into account.

Halo

- *Evaluator assistance.* The number of times the evaluator physically assisted the trainee by holding an incision open, helping to place the vessel loop, etc.
- *Evaluator hint.* At times, the evaluator would suggest an instrument to the trainee or the trainee would ask for advice. For instance, an evaluator may point out scissors to a surgeon who was looking for an instrument, but had made no mention of wanting scissors.
- *Evaluator prompting.* The evaluator had to prompt the trainee to continue speaking or remind the trainee of the goal of the procedure.

3.5 Score Validation

The Maryland global rating score served as our gold standard for performance and will act as our primary outcome measure during the following analyses. In order to ensure its appropriateness as a measure of surgical skill, I first validated this score using the most objective metrics available: task completion, errors, and time. These measures allow me to answer three basic questions that relate to surgical skill: 1) Was the patient saved? 2) How much unnecessary damage was done? and 3) How long did it take? I created a combined measure based on how a hypothetical person would most likely select a surgeon to perform the procedure. The highest priority was whether the surgeon was capable of accomplishing the core objective of the procedure (successfully locate and isolate the artery to stop bleeding before I died from blood loss). Next, the surgeon should not kill the patient in some other fashion while trying to access the artery, as controlling the artery does not help if the patient still dies. Next, the surgeon should avoid lesser, nonfatal errors while saving the patient. Finally, all else being equal, the surgeon should work quickly to minimize total blood loss. However, time is least important within the limit set by exsanguination; working quickly but poorly will not lead to a satisfactory outcome.

Accordingly, I rank-ordered all of the observed surgeries based first on whether the surgeon was able to gain control of the axillary artery proximal to the wound within the specified time limit. This represented a basic yes-no categorization regarding whether the surgeon is capable of performing the procedure. Within these groupings, I then rank-ordered the procedures based on the number of critical (fatal) errors and then less severe errors. This served to order the surgeons based on the damage done in the process of performing the procedure. Finally, I sorted the procedures based on completion time. This sorting process gave me a basic ranking of the surgeons based on how completely, safely, and quickly the task was performed. The best procedure received a rank of 90, and the worst procedure received a rank of one. These rankings were then compared to the Maryland global scores in order to determine how well the global scores corresponded to performance.

I used Spearman's correlations to compare the WSU objective rank-based score against the Maryland global score. The Maryland global ranking demonstrated a significant relationship with the objective rank-based score ($r(88) = 0.76$, $p < 0.01$). I therefore felt confident that the Maryland global score provided a reasonable measure of performance.

3.6 Data Reduction Using Principal Components Analysis

In order to determine whether the WSU variables could account for variance in surgical performance scores, I first performed a data reduction using an exploratory PCA with Varimax rotation. I selected an orthogonal rotation technique in order to produce more interpretable, uncorrelated factors. This technique grouped the WSU variables based on shared variance and allowed me to derive measures of higher-level constructs from our data set.

I used an iterative approach to the PCA. I first entered all of the final WSU variables into the model. The initial results revealed a total of 17 factors with Eigenvalues greater than one. However, examination of the scree plot revealed that only up to 11 factors may have been present in the data. I therefore evaluated models retaining between one and 11 factors in order to determine the best model for the data set. Subsequent examination of the models retained variables loading at least 0.5 on a given factor, with a difference of at least 0.35 between loadings on other factors (mild flexibility of a couple of hundredths in this difference was allowed for conceptually convincing variables). The five-factor model ultimately survived examination and will be described further here. The other models were rejected either because of factors with only one variable, or because of lack of coherence in some factors. Descriptions of the rejected models can be found in Appendix E.

3.6.1 The accepted 5 factor model.

3.6.1.1 5 factor model description.

Rotated factor loadings for the model retaining five factors are listed in Appendix F. Factor 1 in this model included *Expressions of doubt or uncertainty, Backtracking,* and *Realizing a mistake*. Factor 2 consisted of *Time between identifying the artery and placing the vessel loop, Environment adjustment (workspace), Total instrument changes, Instrument changes during the muscle and control phase,* and *Time spent searching for instruments.* Factor 3 was made up of *Proportion of the time using instruments in two* *hands during the incision, Instrument changes during the incision,* and *Shifts between blunt and sharp dissection.* Factor 4 was composed of *Miscellaneous oddities* and *Proportion of the time sharp dissection was used during the muscle phase.* Finally, Factor 5 included *Mentioning things they expect to happen, Naming structures, Knowledge,* and *Balancing constraints.*

Factor 1 contained behaviors described above under goal establishment. Factors 2-5 contained behaviors described under goal enactment, indicating that both of these constructs were captured in my analysis, and that both constructs hang together coherently. Factor 1 appeared to contain variables related to identifying problems. Factor 2 appeared to capture primarily variables related to instrument changes. Factor 3 contained variables related to strategy selection (as switching instruments during the incision typically meant that the surgeon had started using a different dissection method). Factor 4 included behaviors that were not necessarily incorrect, but were not typical of the surgeons as a whole (using a lot of sharp dissection during the muscle phase was not unheard of but was not the typical method of choice). Factor 5 appeared related to declarative or consciously directed behavior. All five factors contained at least two variables and appeared to demonstrate reasonable conceptual coherence, so the model retaining five factors was selected for further investigation.

3.6.1.2 Investigating the 5 factor model.

As noted above, I investigated the model retaining five factors using an iterative process. Because factor loadings depend on the variables included in the model, I removed variables that did not load onto any of the five factors based on our criteria and re-ran the model. I removed variables in this fashion (four total iterations) until all variables in the model loaded onto a factor. The final rotated factor loadings are listed in Table 3.1; intermediate steps are listed in Appendix G.

Table 3.1
Rotated factor loadings for the final iteration of the model retaining five factors.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
	(Instrument	(Strategy)	(Deliberate	(Monitoring)	(Oddities)
Variable	Change)		Behavior)		
Expressions of doubt or uncertainty	-0.11	0.23	0.15	**0.81**	0.00
Realizing a mistake	0.07	-0.19	-0.04	**0.83**	0.17
Time between identifying the artery and placing the vessel loop	**0.71**	0.04	-0.03	-0.24	-0.02
Total instrument changes	**0.90**	0.13	0.11	0.18	-0.01
Instrument changes during the muscle phase	**0.81**	-0.14	-0.10	0.03	0.01
Instrument changes during the control phase	**0.60**	-0.01	0.23	-0.24	0.06
Time spent searching for instruments	**0.80**	0.24	0.10	0.21	0.22
Proportion of the time using instruments in two hands during the incision phase	0.13	**0.84**	0.00	-0.11	0.08
Instrument changes during the incision phase	0.14	**0.82**	0.00	0.13	-0.31
Shifts between blunt and sharp dissection	-0.13	**0.71**	0.10	0.06	0.36
Miscellaneous oddities	0.17	-0.02	0.03	-0.06	**0.81**
Proportion of the time sharp dissection used during the muscle phase	-0.02	0.09	0.01	0.22	**0.76**
Naming structures	0.02	0.27	**0.78**	0.11	0.12
Knowledge	0.02	-0.01	**0.78**	0.14	-0.07
Balancing constraints	0.11	-0.11	**0.67**	-0.13	0.03

Factor 1 in the final model contained *Time between identifying the artery and placing the vessel loop, Total instrument changes, Instrument changes during the muscle and control phase,* and *Time spent searching for instruments.* This factor largely represented behaviors related to instrument changes, with the exception of *Time between identifying the artery and placing the vessel loop*. This apparently stray variable may still be related to instrument changes if the surgeon frequently changed instruments while attempting to dissect around the artery (which would occur during the control phase). Because of this possible connection to changing instruments and the clear relation of the other variables to changing instruments, this factor will hereafter be referred to as the *Instrument Change* factor.

Factor 2 in the final model included *Proportion of the time using instruments in two hands during the incision, Instrument changes during the incision,* and *Shifts between blunt and sharp dissection.* Using instruments in both hands and changing instruments during the incision are likely to represent using an instrument other than a scalpel for the incision, most often scissors. Scissors offer the ability to use sharp dissection (via cutting) as well as blunt dissection (via spreading the tips of the scissors apart to separate tissues). Because of this implied use of a more flexible instrument along with the other variable of switching between blunt and sharp dissection, I believe that this factor represents strategy selection during the procedure. This factor will hereafter be referred to as the *Strategy* factor.

Factor 3 in the final model consisted of *Naming structures, Knowledge,* and *Balancing constraints*. Referring to specific structures or navigating based on anatomical knowledge requires declarative knowledge on the part of the surgeon. Similarly, balancing constraints in the procedure is a deliberate choice based on the values of the surgeon and the larger medical system. Therefore, this factor appears to represent conscious thought processes and will be referred to as the *Deliberate Behavior* factor.

Factor 4 in the final model included *Expressions of doubt or uncertainty* and *Realizing a mistake.* This factor contained variables demonstrating an awareness of the potential for or the occurrence of problems and will be referred to as the *Monitoring* factor for the remainder of the paper. Of particular note is that unlike the diagnosis-related Maryland evaluation items described under goal establishment in section 3.4.1.1, the *monitoring* factor is not separate from the physical procedure; it occurs as the procedure unfolds.

Factor 5 in the final iteration of the model contained *Miscellaneous oddities* and *Proportion of the time sharp dissection was used during the muscle phase.* As sharp dissection was typically only used during the final step of dividing the pectoralis minor muscle, using a high percentage of sharp dissection during the muscle phase is atypical.

This is an acceptable strategy if the surgeon is in a hurry, however. Likewise, *Miscellaneous oddities* represents behaviors that are not necessarily incorrect from the perspective of completing the procedure, but are deviations from generally expected practice such as an especially large or small incision, or beginning the incision in the armpit rather than in the chest. As such, this factor is termed the *Oddities* factor.

3.6.2 Computing factor scores.

I computed factor scores to facilitate further analysis. I first converted each of the variables within the factors to z-scores. These z-scores were then averaged across the variables within each of the final five factors to form factor scores for each factor. These scores served as independent measures for further analyses to account for variance in performance scores.

Table 3.2 summarizes the Maryland and WSU factor contributors to goal establishment and goal enactment, along with the source of the variables in each (Maryland evaluation, think-aloud protocol, or video protocol). Figure 3.1 illustrates the overlap among the contributing components to the Maryland and WSU variables within goal enactment (a similar effort indicated no overlap within goal establishment).

Table 3.2

Maryland and WSU factor contributors to goal establishment and goal enactment.

Construct	Variable	Source
Goal Establishment	Q1	Evaluation
	Q3	Evaluation
	Q12	Evaluation
	Monitoring	Think-aloud
Goal Enactment	Q7	Evaluation
	Q8S1	Evaluation
	Q8S2	Evaluation
	Q9	Evaluation
	Instrument changes	Video
	Strategy	Video
	Oddities	Video
	Deliberate behavior	Think-aloud

Note: Maryland variables are enumerated, while the WSU variables are named based on my interpretation.

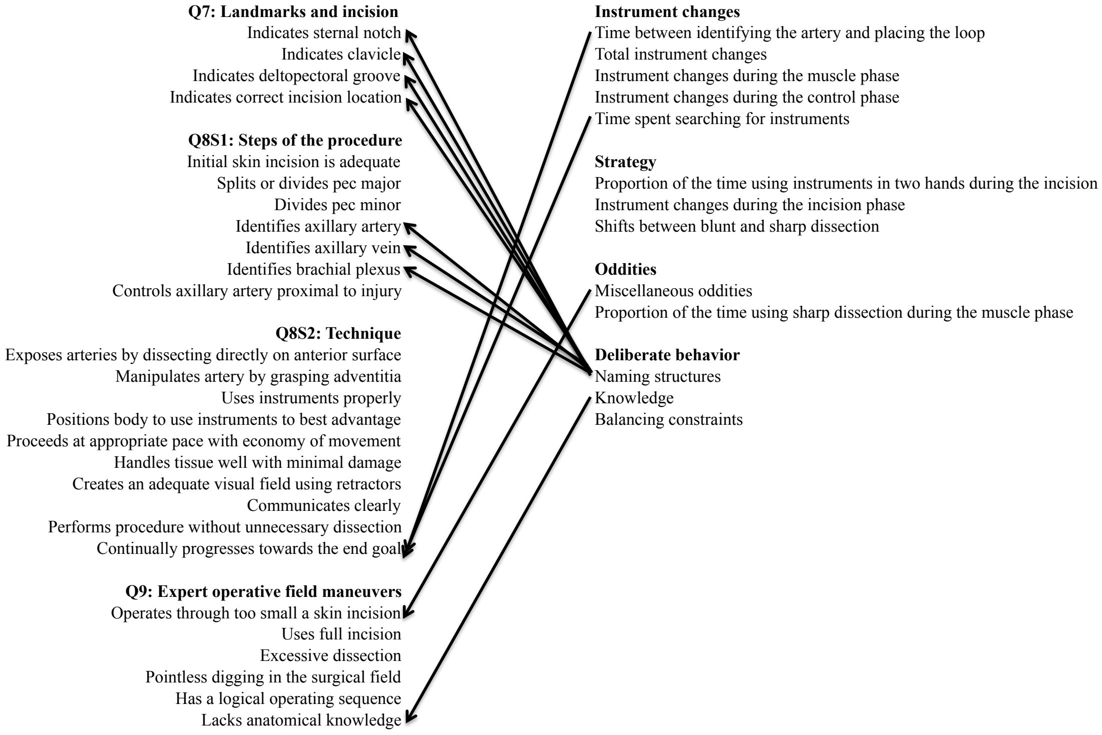

Q7: Landmarks and incision
Indicates sternal notch
Indicates clavicle
Indicates deltopectoral groove
Indicates correct incision location
Q8S1: Steps of the procedure
Initial skin incision is adequate
Splits or divides pec major
Divides pec minor
Identifies axillary artery
Identifies axillary vein
Identifies brachial plexus
Controls axillary artery proximal to injury
Q8S2: Technique
Exposes arteries by dissecting directly on anterior surface
Manipulates artery by grasping adventitia
Uses instruments properly
Positions body to use instruments to best advantage
Proceeds at appropriate pace with economy of movement
Handles tissue well with minimal damage
Creates an adequate visual field using retractors
Communicates clearly
Performs procedure without unnecessary dissection
Continually progresses towards the end goal
Q9: Expert operative field maneuvers
Operates through too small a skin incision
Uses full incision
Excessive dissection
Pointless digging in the surgical field
Has a logical operating sequence
Lacks anatomical knowledge
Instrument changes
Time between identifying the artery and placing the loop
Total instrument changes
Instrument changes during the muscle phase
Instrument changes during the control phase
Time spent searching for instruments
Strategy
Proportion of the time using instruments in two hands during the incision
Instrument changes during the incision phase
Shifts between blunt and sharp dissection
Oddities
Miscellaneous oddities
Proportion of the time using sharp dissection during the muscle phase
Deliberate behavior
Naming structures
Knowledge
Balancing constraints

Figure 3.1. Overlap between Maryland and WSU variables within goal enactment. *On the surface, “uses instruments properly” appears related to the instrument change items. However, “uses instruments properly” refers to actual use (i.e., holding the instrument correctly or avoiding backhanded use).

Aside from the obvious relation of “naming structures” to several Maryland evaluation items, the majority of the identified WSU variables are separate from the Maryland items. Goal establishment constructs showed no overlap between WSU and Maryland variables with this examination, and relatively little overlap was seen within goal enactment in Figure 3.1 above. On the aggregate, the WSU factors and the Maryland evaluation items appear to capture distinct constructs. I identified no clear one-to-one mapping between any of my identified WSU factors and the grouped Maryland items, although variables contributing to the deliberate behavior factor did overlap with several Maryland variables. This potential overlap is addressed as part of later analyses (particularly section 4.2.1.1 in Chapter 4). Overall, I felt comfortable that any relationships observed between the WSU and Maryland variables would not be due to the fact that I have simply recoded the same things evaluated in the Maryland study.

3.7 Examining the Benefit of Content vs Simple Word Count

As will be discussed in later chapters (particularly Chapter 6), many of the best surgeons completed the procedure very quickly. The relative speed of these surgeons caused shorter transcripts. A simple word count of the transcripts may therefore serve as a useful predictor and preclude the need for further analysis. In order to establish that a content analysis of the transcripts provides a useful contribution to the data set, I examined whether the WSU factors accounted for variance in Maryland global scores beyond that accounted for by word count alone. I ran a stepwise regression predicting Maryland global scores using word count in the first step and the five WSU factor scores in the second step. **Model 3.1** indicated that word count alone significantly predicted global outcome scores ($R^2 = 0.14$, $F(1,67) = 11.15$, $p < 0.01$). **Model 3.2** with the five WSU variables included also predicted global scores ($R^2 = 0.47$, $F(6,62) = 9.32$, $p <$

0.01). The change in R^2 was statistically significant (F change (5,62) = 7.82, $p < 0.01$), indicating that the content of the transcripts in the form of the WSU factors contributed significant explanatory power over word count alone.

My analyses have identified five new predictors derived from variables in my WSU coding scheme. Drawn from both think-aloud and video-based data, I have contributed one new predictor to goal establishment (that is better integrated into skilled behavior rather than kept separate) and four new predictors to goal enactment. These factors contribute additional explanatory power in accounting for variance in global outcome scores (i.e., content of speech matters more than the amount in predicting performance). These WSU factors represent generalized higher-level processes, particularly the monitoring factor. They capture cognitive processes, particularly related to monitoring – the "why" to the Maryland variables' "what". In Chapter 4 I utilize these new predictors to capture variance in global performance scores and demonstrate the importance of establishment and enactment to skilled behavior, particularly the importance of deliberate processes such as monitoring.

Chapter 4 – Sampling Expertise in Performance Measures

4.1 Introduction

The definition of expertise establishes parameters for what constitutes expert performance, with implications for the behaviors captured in evaluations. Improving theory will therefore improve methods for evaluating performance by allowing us to properly sample the domain to capture the behaviors that specify expertise. How a domain is sampled affects how items are weighted, which in turn affects performance scores. Performance evaluation tools must sample domain-relevant skills and behaviors in a principled fashion in order to ensure content validity. While a content analysis provides one dimension of the sampling problem (for example, different kinds of vascular surgeries), a cognitive analysis provides a complementary dimension. For example, a popular cognitive analysis distinguishes between declarative and procedural knowledge (Anderson, Matessa, & Lebiere, 1997). Accordingly, one might design a performance measurement tool to sample both types of knowledge. However, the domain focus of content analyses is too specific to generalize across tasks, and the knowledge focus of cognitive analyses can be too broad to give a clear picture of the task in the context of the environment. Neither approach fully addresses how people actually identify and solve problems in the work environment (i.e., the function of an agent's knowledge rather than its organization).

Here I explore an alternative distinction, between goal establishment and goal enactment processes. Rather than serve as causal constructs, the concepts of goal establishment and goal enactment help describe how the agent interacts with the environment over time to identify problems and implement solutions in the world. I argue that both types of behavior must be sampled in order to ensure that performance scores accurately reflect the components of skilled behavior.

4.1.1 WSU and Maryland predictors linked to goal establishment and goal enactment.

As described in Chapter 3, the identified WSU factors encompass both goal establishment and goal enactment. Goal establishment is represented by the *monitoring* factor, composed of behaviors related to evaluating how well the procedure is progressing or is likely to progress. Goal enactment is represented by the *instrument change, strategy, deliberate behavior,* and *oddities* factors. Each of the goal enactment factors contain behaviors related to executing the procedure such as what tools to use, how best to use them, prioritizing values in selecting a course of action, or staying oriented in the body. Though many of the Maryland evaluation items also address aspects of goal establishment and goal enactment (e.g., making a diagnosis, holding instruments correctly, etc.), the identified WSU factors address these functions in a more generalized manner by emphasizing broader cognitive activity rather than physical execution. I now use these identified WSU factors along with the Maryland predictors to examine variance in Maryland global scores, making the case that the WSU factors represent higher-order constructs that will prove useful for guiding sampling of domain-relevant behaviors.

4.1.2 The current analysis.

I sought to demonstrate the relevance and generalizability of goal establishment and goal enactment behaviors in measuring skilled skilled performance by using them to capture variance in the Maryland global scores. Below I establish the relationship between process variables and the global outcome score, considering both the Maryland and WSU process variables in several steps. To make the argument that the WSU variables represent higher-order constructs relative to the Maryland variables, I first employ WSU variables as predictors both as a single group and controlling for Maryland variables in multiple regression analyses to show variance in Maryland global scores accounted for by the WSU variables. Second, I use mediation analysis to demonstrate that the WSU variables identify higher order constructs in the Maryland variables. Finally, I provide convergent evidence regarding the general relevance of these constructs by predicting a broader surgical skill performance score (TRI) that spans multiple vascular procedures as well as an outcome score from a completely separate procedure.

Although the reader will recall from Chapter 2 that the residents' pre- and post-ASSET procedures were completed within-subjects, I have elected to ignore this aspect of the data set for the current analysis in order to include the attending surgeons' data as well. Including the attending surgeons' procedures helps to broaden the range of observed behaviors and scores and better preserves the expert-novice comparison paradigm. The potential ramifications of ignoring the within-subjects nature of the study are discussed further in later chapters. Including the attending surgeons also precludes the possibility of utilizing multilevel techniques (procedures nested within surgeons) because the attending surgeons only have one procedure each. These analyses therefore rely on standard regression techniques.

4.2 Results

4.2.1 Predicting Maryland global scores using WSU and Maryland variables.[4,5]

I sought to examine the proportion of variance in Maryland global scores accounted for by the Maryland and WSU variables. P-P plots for the analysis indicated that the assumption of normality in the residuals was violated in my data set for the Maryland global scores. However, regression is considered to be robust to violations of this assumption (Cohen, Cohen, West, & Aiken, 2003) and transforming our data would have complicated interpretation of the results. Because of the robust nature of regression and the exploratory nature of our analyses I therefore elected to continue with the regression analyses without transforming the data.

Further, the reader will recall that the Maryland study data consisted of pre- and post-ASSET testing of residents, along with a separate group of attending surgeons. This testing arrangement produced some comparisons that were within-subjects, some that were between subjects, and some that were both. In comparisons using pre-ASSET, post-ASSET, and attending surgeon procedures, I elected to ignore the repeated measures aspect of the study design. Though this decision likely decreases power overall by adding variance to the error term, I made this tradeoff in order to include the attending surgeons and thus broaden the potential observed range of both predictor and outcome scores.

[4] Analyses for the Maryland global scores were also performed on the Maryland IPS scores and WSU objective rank scores. The results showed slight differences but followed the same general pattern of effects. The IPS and objective rank score analyses can be found in Appendices H and I, respectively, identified by the same section headings as used in the document.

[5] Regression models predicting Maryland global scores were run using two methods: with all variables included (reported in the document) and after allowing nonsignificant predictors to drop out (reported in Appendix J). Both methods were used in order to confirm that our results were not due to extraneous variables in the predictor data; the observed patterns of results were the same across methods.

4.2.1.1 Variance accounted for by WSU variables alone.

To establish whether the identified WSU constructs were indeed relevant to performance, I first predicted the Maryland global scores using the WSU variables in a series of regression models. I entered all five WSU variables (scores on the instrument change factor, strategy factor, deliberate behavior factor, monitoring factor, and oddities factor) into regression models for the Maryland global outcome measure. Overall, WSU variables accounted for significant variance in global scores (**Model 4.1;** $R^2 = 0.47$, $F(5, 63) = 11.16$, $p < 0.01$). Specific scores on the goal enactment predictors *instrument change* factor ($\beta = -0.22$, $t(63) = -2.33$, $p = 0.02$) and *strategy* factor ($\beta = -0.26$, $t(63) = -2.80$, $p = 0.01$), as well as the goal establishment predictor *monitoring* factor ($\beta = -0.54$, $t(63) = -5.74$, $p < 0.01$) all significantly predicted global outcome scores. Each of the WSU predictors (representing both goal establishment and goal enactment) was negatively related to Maryland global performance scores, indicating that better-performing surgeons displayed fewer *instrument change, strategy,* and *monitoring* behaviors. Scores on the goal enactment factors *deliberate behavior* and *oddities* were not significant ($p = 0.20$ and $p = 0.68$, respectively). The absence of significant effects for deliberate behavior dampens any concern regarding a simple overlap between Maryland and WSU variables in Figure 3.1 from Chapter 3.

4.2.1.2 Variance accounted for by WSU variables, controlling for Maryland variables.

To examine whether the new constructs provide useful additional information, I investigated the proportion of variance accounted for in global scores by the WSU and Maryland variables together. I ran a series of two-stage linear regression models entering the Maryland variables in the first step and the WSU variables in the second step. I selected the final set of Maryland variables aggregated from the evaluation form (described in section 3.4.1): Q1 (suspected injury), Q3 (additional studies), Q7 (landmarks and incision), Q8S1 (steps of the procedure), Q8S2 (technique), Q9 (expert operative field maneuvers), and Q12 (pitfalls). These variables were selected based on their combination of reliability and inclusion in the Maryland IPS score (section 3.4.1.3), indicating their perceived unique contribution to performance by the surgeons (as well as facilitating comparisons across analyses using various global and IPS measures). WSU variables entered included the five retained factor scores (instrument change score, strategy score, deliberate behavior score, monitoring score, and oddities score). I examined the proportion of variance accounted for by each model as well as the change in R^2 between models.

Model 4.2 indicated that the Maryland variables alone significantly predicted global outcome scores (R^2 = 0.78, F(7, 60) = 29.58, p < 0.01). **Model 4.3** with the five additional WSU variables included also predicted global scores (R^2 = 0.82, F(12, 55) = 20.20, p < 0.01). This 4% change in R^2 was just over the cutoff for statistical significance (F change (5,55) = 2.36, p = 0.052).

Within model 4.2, only the Maryland goal enactment predictors Q8S1 (steps of the procedure; β = 0.54, t(60) = 6.50, p < 0.01) and Q8S2 (technique; β = 0.52, t(60) = 6.80, p < 0.01) significantly predicted global scores. These findings indicate that higher-performing surgeons follow the outlined steps of the procedure more closely and demonstrate better operative technique. Within model 4.3 with the WSU variables included, Q8S1 and Q8S2 remained significant predictors while scores on the *deliberate behavior* factor were also a significant predictor of global scores (Q8S1 β = 0.44, t(55) = 4.98, p < 0.01; Q8S2 β = 0.47, t(55) = 5.97, p < 0.01; *deliberate behavior* β = -0.17, t(55)

= -2.62, *p* = 0.01). I note that the additional contribution of the *deliberate behavior* factor is somewhat unexpected given that this factor was not a significant predictor in the WSU-variable-only analysis of section 4.2.1.1. This finding is most likely a statistical artifact wherein the Maryland predictors captured enough variance in global outcome scores to reduce the standard error of the *deliberate behavior* factor, allowing it to become significant (LaHuis, personal communication). It appears that the WSU variables do not account for additional variance in Maryland global outcome scores beyond that accounted for by the Maryland variables, though the reduction in the predictive value of the WSU variables after controlling for Maryland variables hints at possible mediation effects. I examined this possibility below.

4.2.2 Examining mediation between WSU and Maryland variables.

The failure of the *instrument change, strategy, and monitoring* factors to account for variance in the full model for Maryland global scores suggests the possibility of shared variance between the Maryland evaluation items and the WSU factor scores. I reasoned that the WSU variables likely represented higher-order cognitive constructs, while scores on the Maryland evaluation items represented the task-specific manifestation of these constructs. I therefore examined the possibility of mediation between the Maryland and WSU variables (Figure 4.1). Mediation would indicate that the WSU variables not only add value to the current rating system, but capture higher-order constructs that facilitate more generalized measures of skill.

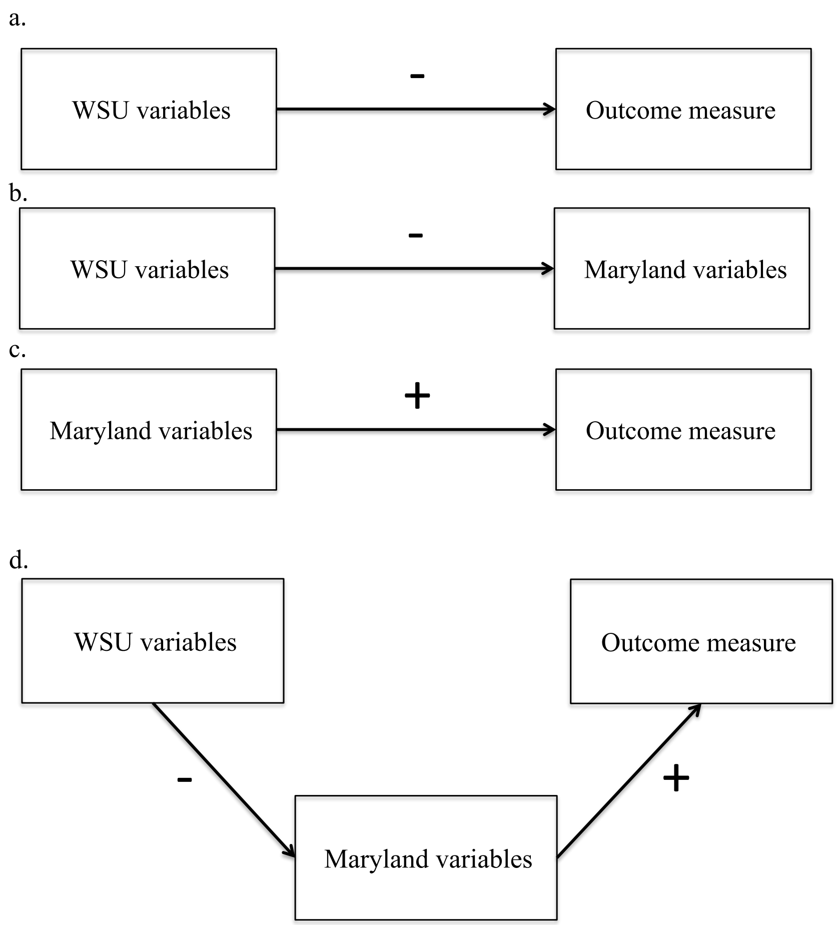


Figure 4.1. Mediation between WSU factors and global outcome scores. (a) Relationship between WSU predictors and outcome measures. (b) Relationship between WSU predictors and Maryland predictors. (c) Relationship between Maryland predictors and outcome measures. (d) Form of the general relationship between WSU, Maryland, and outcome measures. The WSU variables are negatively related to the Maryland variables, which in turn are positively related to the outcome measures. Increases in the WSU variables are associated with decreases in the Maryland variables, which are in turn associated with decreases in outcome scores.

Four criteria must be met to establish mediation (Baron & Kenny, 1986). First, the causal variable (the WSU variables in this case) should be related to the outcome (Figure 4.1a. Next, the causal variable should be related to the mediating variable (the Maryland variables; Figure 4.1b). Third, the mediator variable must be related to the outcome variable (controlling for the causal variable; Figure 4.1c). Finally, the relationship between the causal variable and the outcome variable must be reduced or even eliminated when controlling for the mediating variable. Criterion one was established in model 4.1, where scores on the instrument change, strategy, and monitoring factors accounted for variance in global scores. Criteria three and four were established in model 4.3, where the Maryland variables Q8S1 and Q8S2 significantly predicted global scores in the presence of the WSU variables, and scores on the instrument change, strategy, and monitoring factors were no longer significant predictors of global scores when controlling for the Maryland variables. I examined criterion two (the relationship between causal variables and mediating variables) by correlating the WSU and Maryland variables (Table 4.1).

Table 4.1
Correlations between WSU factor scores and Maryland evaluation items.

Variable	Monitoring factor (goal establishment)	Instrument change factor (goal enactment)	Strategy factor (goal enactment)	Deliberate behavior factor (goal enactment)	Oddities factor (goal enactment)
Goal establishment					
Q1 (suspected injury)	0.00 (88)	0.18 (67)	-0.19 (88)	0.11 (88)	-0.01 (72)
Q3 (additional studies)	0.12 (88)	0.02 (67)	-0.01 (88)	0.01 (88)	0.02 (72)
Q12 (pitfalls)	**-0.25 (88)**	0.08 (67)	0.07 (88)	0.15 (88)	0.01 (72)
Goal enactment					
Q7 (landmarks and incision)	**-0.30 (88)**	-0.04 (67)	**-0.27 (88)**	0.05 (88)	-0.20 (72)
Q8S1 (steps of the procedure)	**-0.46 (88)**	**-0.24 (67)**	**-0.39 (88)**	0.04 (88)	**-0.25 (72)**
Q8S2 (technique)	**-0.39 (88)**	**-0.27 (67)**	**-0.34 (88)**	0.12 (88)	-0.06 (72)
Q9 (expert operative field maneuvers)	**-0.31 (87)**	-0.08 (66)	-0.18 (87)	0.20 (87)	0.05 (71)

Note: Correlations in bold are significant at $p < 0.05$.

All of the WSU factors (with the exception of the deliberate behavior factor) correlated significantly with at least one of the Maryland evaluation items, including items Q8S1 and Q8S2, which predicted global scores. Further, both the WSU and Maryland variables were correlated with outcome scores and the direction of the correlations indicated that mediation was plausible (Table 4.2). I therefore concluded that the Maryland assessment items mediated the relationship between the WSU factor scores and global outcome scores (as illustrated in Figure 4.1 above). Based on the directions of the relationships between WSU and Maryland predictors with outcome measures, the direction of correlations between WSU and Maryland variables, and the inability of WSU variables to account for variance in outcome measures beyond that accounted for by the Maryland predictors, the Maryland variables completely mediated the relationship between the *instrument change, strategy,* and *monitoring* WSU factors and Maryland global scores. The analyses indicate that higher scores on the WSU factors are associated with lower scores on the Maryland variables, which are in turn associated with lower outcome scores. The hypothesized reason for the negative relationship between WSU variables and Maryland variables (and therefore outcome scores) will be explored in the discussion section. This mediation suggests that the WSU variables represent higher-level constructs that may generalize to other procedures.

Table 4.2
Correlations of Maryland and WSU variables with the Maryland global scores.

Variable	Global scores
Goal establishment	
Q1 (suspected injury)	0.14 (88)
Q3 (additional studies)	-0.14 (88)
Q12 (pitfalls)	0.17 (88)
Monitoring factor	**-0.48 (88)**
Goal enactment	
Q7 (landmarks and incision)	**0.51 (88)**
Q8S1 (steps of the procedure)	**0.85 (88)**
Q8S2 (technique)	**0.82 (88)**
Q9 (expert operative field maneuvers)	**0.47 (87)**
Instrument change factor	**-0.26 (67)**
Strategy factor	**-0.42 (88)**
Deliberate behavior factor	-0.01 (88)
Oddities factor	-0.18 (72)

Note: Correlations in bold are significant at $p < 0.05$.

4.2.3 Converging evidence of higher-level constructs via TRI analysis.

As an additional check to see whether the identified WSU variables indeed represent higher level constructs that can account for skill across procedures, I examined whether my identified constructs accounted for variance in the Maryland TRI measure. Recall that the TRI measure is calculated based on the evaluations of all four procedures included in the parent Maryland study. I generated an adjusted TRI measure excluding the axillary artery procedure to create a measure of performance across the three remaining procedures not used to generate the WSU items.

I compared the ability of the WSU variables and Maryland variables to account for variance in the adjusted TRI measure (excluding the axillary artery exposure). When predicting the adjusted TRI measure, the Maryland variables fared quite well (**Model 4.4**; $R^2 = 0.66$, $F(7,80) = 22.03$, $p < 0.01$). Interestingly, only Q8S2 (technique points) significantly predicted the adjusted TRI score when controlling for the other Maryland variables ($\beta = 0.62$, $t(82) = 6.77$, $p < 0.01$). The WSU variables also predicted the adjusted TRI score, though not quite as well as the WSU variables (**Model 4.5**; $R^2 = 0.24$, $F(5, 62) = 3.92$, $p < 0.01$). Only the score for the monitoring factor was a significant predictor in the presence of the other WSU variables ($\beta = -0.44$, $t(84) = -3.87$, $p < 0.01$). To see whether the WSU variables added any additional explanatory power over the Maryland variables, I generated a stepwise model (**Model 4.6**) predicting the adjusted TRI with the Maryland variables in the first step and the WSU variables in the second step. The analysis indicated that adding the WSU variables in addition to the Maryland variables did not add significant explanatory power to the model (R^2 change = 0.06, F change $(5,54) = 2.03$, $p = 0.09$).

Because two of the three remaining procedures included in the adjusted TRI were highly similar to the axillary procedure (all were vascular procedures requiring identification and control of a major blood vessel), I also examined the ability of the WSU and Maryland variables to predict the IPS score for the fasciotomy procedure. This procedure does not require the same type of handling of arteries or other tissue and may serve as a better gauge of the generalizability of the Maryland and WSU variables. The model predicting the fasciotomy IPS again showed good predictive ability for the Maryland variables (**Model 4.7**; $R^2 = 0.51$, $F(7, 80) = 12.08$, $p < 0.01$). The Maryland variables Q8S1 ($\beta = 0.33$, $t(82) = 2.59$, $p = 0.01$) and Q8S2 ($\beta = 0.33$, $t(82) = 3.07$, $p < 0.01$) were significant in this model. The model predicting the fasciotomy IPS using the WSU factors was also significant, though not as successful as the Maryland variables (**Model 4.8**; $R^2 = 0.34$, $F(5, 62) = 6.28$, $p < 0.01$). Once again, a stepwise model (**Model 4.9**) predicting the fasciotomy IPS with the Maryland variables in the first step and the WSU variables in the second step indicated that adding the WSU variables in addition to the Maryland variables did not add significant explanatory power to the model (R^2 change = 0.07, F change (5,54) = 2.14, $p = 0.07$).

4.3 Discussion

4.3.1 Summary of findings.

In this chapter I have demonstrated that identified WSU constructs related to goal establishment and goal enactment are useful to explain variance in Maryland global outcome scores. Mediation analyses indicated that the relationship between WSU variables and outcome scores is mediated by the Maryland variables, implying that the WSU variables capture more generalized constructs reflected in the criteria used in the Maryland evaluation form. Taken together, these analyses indicate that goal establishment and goal enactment are useful constructs for capturing skilled performance. This conclusion is tentatively supported by the ability of the WSU variables to account for the more general Maryland outcome measures such as the adjusted TRI and fasciotomy IPS scores. These findings extend the work of my thesis by linking goal establishment and goal enactment to performance, and speak to the importance of goal establishment and goal enactment in guiding performance measures, as well as the interaction between goal establishment and goal enactment in the context of skilled performance.

4.3.2 Incorporating the expertise literature.

The relationship between my WSU variables and the Maryland global outcome score both affirms existing beliefs on expertise and illustrates the need to expand how expertise is conceptualized. Cognitive theory describes experts as having large stores of knowledge organized in a high-level manner, retrieved and employed via associative processes. Skill-based psychomotor tasks do not require representational cognition on the part of the expert (Rasmussen, Pejtersen, & Goodstein, 1994). My analyses indicating lower scores on the *instrument change, strategy*, and *monitoring* factors for experts (section 4.2.1.1) support the notion that the best surgeons executed the task smoothly, with little need for corrective action once a plan was enacted. However, the importance of the *monitoring* factor hints that skilled performance involves cognitively penetrable factors in addition to associative processes (addressed further in Chapters 5 & 6). All of these processes must be captured in order to best measure skilled performance.

Whereas my thesis identified the constructs of goal establishment and goal enactment, I lacked standardized tasks and outcome measures to link these constructs to performance. In this study, three of my identified factors encompassing both goal establishment (*monitoring*) and goal enactment (*instrument changes* and *strategy*) were related to global performance scores, mediated by the procedure-specific items used in the Maryland trauma study. These constructs reflect the nature of expertise, as described in the cognitive literature (e.g., the ability to devote resources to self-monitoring as described by Beilock et al., 2004; MacIntyre et al., 2014; and McPherson, 2000). The identified factors demonstrated negative relationships with global outcome scores, due in large part to the nature of surgical expertise described in Chapter 1.

In the case of the *Instrument Change* and *Strategy* factors, recall from Chapter 1 that expert surgeons are able to anticipate what instruments they will need and that experts do not rely on interaction with the world for feedback (i.e., they know what action to take and how to avoid error). The finding that experts would change instruments or alter their dissection strategy less frequently than less-skilled surgeons makes sense in this context. In the case of the *Monitoring* factor, the negative relationship with outcome scores is likewise due to the nature of expertise (discussed further in sections 8.1.2.4 and 8.3.4). As discussed in Chapter 1, novices sometimes rely on interaction with and negative feedback from the world for information whereas experts learn to avoid negative consequences in the world without requiring such feedback. Experts are less likely to display the behaviors articulated within the *monitoring* factor by definition (i.e., experts will make fewer mistakes). Such alignment with expertise theory further underscores the utility of the identified WSU factors in capturing expert performance and bolsters the claim that these constructs should be used to guide behavioral sampling. The fact that these findings align with previous research on surgical expertise (e.g., Tien et al., 2015) also increases confidence that I have identified meaningful factors and that the relationship between these factors and performance is not spurious.

4.3.3 The interaction between goal establishment and goal enactment.

Based on the direction of the relationships between WSU factors and outcome scores in section 4.2.1.1 above, experts appear to display fewer goal establishment (monitoring) behaviors during the procedures (possibly because problems are less likely to arise for experts during the procedure). The best surgeons also appear to display smoother goal enactment than less skilled surgeons in the form of fewer instrument changes and strategy shifts, perhaps even facilitating their own success (for example, proper instrument selection facilitates cleaner technique, which may aid in identifying anatomy or navigating through the body).

In this way, goal establishment and goal enactment interact. Good goal establishment (e.g., correct diagnosis) helps to set the surgeon on the correct path to facilitate successful goal enactment. Good goal enactment behaviors (e.g., clean technique and properly executed steps as evidenced in the positive relationship between the Maryland variables Q8S1 and Q8S2 and Maryland global outcome scores in section 4.2.1.2) help facilitate navigation through the body and reduce the likelihood of unexpected problems, limiting the need for new goal establishment. Though goal establishment typically precedes goal enactment in time, goal enactment can help generate new goals. Goal establishment and goal enactment shape and are shaped by the surgeons' continual interaction with the world.

4.3.4 Sampling domain behaviors.

Assessment items based on single tasks are time-consuming to develop, and only provide insight into the particular task under examination. Measures derived from content-based analyses such as the Maryland predictors are too domain-focused to generalize beyond the original task. On the other hand, cognitive analyses tend to neglect the environment in favor of the agent's knowledge. My findings offer suggestions of how best to capture expertise within a domain by providing insight into the factors that contribute to expert performance. Goal establishment and goal enactment incorporate both the surgeons' knowledge and the work context to describe how experts interact with the environment over time to identify problems (goal establishment) and implement solutions (goal enactment). Although goal establishment and goal enactment likely share similar cognitive foundations, their functional distinction is useful in order to guide behavioral sampling to ensure that performance metrics capture expertise more completely.

Predictably, the task-specific Maryland variables used to measure performance in the parent Maryland training study did quite well in predicting total variance in outcome scores, and outperformed the WSU factors in predicting both the adjusted TRI measure and the fasciotomy-specific IPS measure. However, these predictors focus on task-specific actions, neglecting the role of cognition. Although on the surface goal establishment (in the form of diagnosis and testing) and goal enactment (in the form of the procedural items) are included in the Maryland evaluation, several individual items fared poorly at capturing performance. Many of the Maryland variables failed to account for variance in global scores, indicating that the evaluators did not consider diagnosis, testing, or even certain procedural actions in their global scoring. In contrast, both goal establishment and goal enactment as operationalized in my analyses captured variance in performance scores, indicating that I have identified useful constructs that help operationalize task-specific behaviors in a way more applicable to performance measurement. In contrast to a priori goal establishment of the Maryland evaluation (diagnosis), my WSU *monitoring* factor addresses goal establishment in the midst of task execution, integrating goal establishment rather than keeping it separate from goal enactment.

I have identified higher-level constructs related to general surgical skill that may facilitate a better view of a surgeon's overall skill level, rather than their skill on any particular type of procedure, evidenced by the ability of my WSU constructs to capture variance in cross-procedure performance measures (the adjusted TRI score and fasciotomy IPS score) as well as global scores on the particular procedure examined here. Although the Maryland items were stronger predictors, I believe that this is partly due to shared criteria across the procedures. Recall that the TRI and IPS scores were calculated exclusively using scores on the procedure-specific Maryland evaluation items across the four procedures studied during the parent Maryland training study; TRI and IPS scores would therefore be expected to correlate strongly with these items. When predicting both the adjusted TRI and fasciotomy IPS score with the Maryland items, only the Maryland items related to technique significantly predicted outcome scores. Based on the results of Models 4.4 and 4.7, technique (holding instruments correctly, etc.) appears to be the most critical factor in scoring, which would not be expected to change across procedures. Upon closer inspection of the evaluation forms across each procedure, the items used to evaluate technique are nearly identical across all three procedures. These shared Maryland variables across procedures would therefore be expected to perform quite well in predicting variance in TRI and IPS scores, regardless of the procedure in question.

Although my identified WSU factors did not add explanatory power beyond the Maryland items (specifically the technique items), I still believe in the generalizability of these factors and their utility in guiding behavioral sampling for performance measures. The WSU factors were able to predict variance in global scores, a generalized TRI measure, and an IPS score for a completely separate type of procedure. Unlike the Maryland variables that had the advantage of shared evaluation items across procedures, my WSU factors are completely separate from the evaluation forms. Capturing any variance at all in other procedures is an encouraging sign for the generalizability of these measures. Further, only technique appears to be of any interest in the Maryland items. Surgical expertise clearly involves more than how well one can cut. My WSU factors move beyond raw technique to capture how surgeons interact with the world in real time to monitor their behaviors. Such constructs should serve to guide the selection of items for assessment tests. By assessing these behaviors, perhaps we can begin to create performance assessment tools that better capture domain-level skill without the need to administer multiple task-specific assessments, or refine the development of task-specific assessments to better account for performance.

4.3.5 Conclusions.

We often utilize task-specific performance measures in the belief that if we want to know how good a person is at something, we should measure their skill at that particular thing. However, domains and careers are not made up of single tasks. Surgeons must have knowledge of a broad range of procedures, not just one. We should strive to assess “how prepared is this person to be a surgeon?”, not only “how prepared is this person to execute procedure X?”. Goal establishment and goal enactment facilitate such assessment by offering a path to more generalizable performance measures as well as incorporating existing knowledge regarding the nature of expertise.

Chapter 5 – Self-awareness and Expert Performance

5.1 Introduction

No two problems are ever identical in the open, real world. In fact, several lines of research emphasize the need to accommodate variability between problems. One of the most important Gestalt psychologists, Wertheimer (1959), asserts that mastery requires the ability to act across multiple contexts. Rasmussen (1994) acknowledges this with the inclusion of knowledge-based behavior in his SRK framework, and Vicente's (1999) decision ladder also incorporates knowledge-based reasoning when associative processes fail. The situated cognition perspective (Greeno & Moore, 1993) makes adaptive capability the central theme.

5.1.1 Cognitive penetrability and adaptation.

Persisting debate centers around the cognitive functions required for this adaptive capability. While the typical proponent of situated cognition eschews mentalism (e.g., Suchman), psychology has a long history of promoting reflective capability as central to our intelligence. For example, the equally important Gestalt psychologist, Luchins (1942) is famous for trapping thoughtless repetition (functional fixity) in his water jugs task. More recently, the deliberate practice framework (Ericsson, Krampe, & Tesch-Romer, 1993) emphasizes the role of conscious effort and reflection in learning. The relevance of the monitoring factor to performance highlighted in Chapter 4 suggests the *cognitive penetrability* of skilled performance[6]. Pylyshyn first suggested the concept of cognitive penetrability, which refers to the ability of goals and beliefs to influence action. Cognitive penetrability captures the long-standing concern for the influence of top-down, intentional and/or semantic context on the conduct of bottom-up, feature driven automatic processes.

Whereas Ericsson focused primarily on tasks such as music, chess, and athletic performance, surgery does not offer the same opportunities for practice. The wide variety of procedures a surgeon may be asked to perform precludes practice and mastery of all of them during training, demanding a level of learning and reflection during performance. Safety considerations further require surgeons to be aware of their own skills, and adjust their self-perception in response to feedback from the world.

The experimental work addressing this issue generally manipulates the presence or absence of context to demonstrate the influence on task performance (Kaakinen, Hyona, & Viljanen, 2011). When context is not controlled, the burden of demonstration shifts to complementary dependent measures. The conceptualization of task execution as cognitively penetrable gains support through correlations of performance with measures of self-awareness. In this chapter I further explore the cognitive penetrability of skilled behavior via the self-awareness demonstrated by the surgeons. Below I provide data concerning the relationship between self-awareness and performance, relative to both goal establishment and goal enactment.

[6] Although the monitoring factor demonstrated a negative relationship with performance, sampling issues (of both the surgeons and behavior) alluded to at the end of Chapter 1 and discussed in Chapter 8 likely contributed to this finding.

5.1.2 The current analysis.

I examined the surgeons' awareness of their own skill by analyzing a) the relationship between surgeons' a priori confidence and performance, b) whether confidence changed in response to information from the world, and c) the influence of experience and skill on the predictive value of self-confidence ratings. Such analyses serve to establish the extent to which surgeons appear able to judge their own ability (and hence the extent to which the surgeons may anticipate that additional monitoring behaviors may be necessary). Later adjustment of confidence in response to performance indicates not only an awareness of outcomes, but also the ability of feedback from the world to influence conscious perceptions of ability (and not merely automatic execution of procedural knowledge). In addition, surgeons must be able to monitor their performance regardless of experience or skill level. An ability on the part of novice or poorly performing surgeons to judge skill increases the plausibility that such awareness can be used to increase one's self-monitoring behaviors. I show that surgeons' self-confidence before procedures does predict performance, indicating that the surgeons seem to be aware of their own capability. Further, self-confidence in ability to perform the procedure (but not self-confidence in one's anatomical knowledge) changed in response to performance, partially indicating that the surgeons are aware of their performance and incorporate feedback into their perceptions of their ability. These findings were consistent regardless of surgeons' experience or performance.

I examined the relationship between confidence and performance within two areas: anatomical knowledge and overall performance. I first establish a relationship between surgeons' predictions of performance and global scores, and demonstrate that the surgeons' self-confidence and performance appear to be driven by the same variables.

These analyses demonstrate that surgeons appear to be aware of their own performance generally, and the elements of behavior that drive performance. I next demonstrate that confidence changes in response to training and past performance, further indicating that surgeons are aware of their skill. Finally, I examine whether self-awareness changes as surgeons gain experience or skill. These final analyses examining the relationship between self-awareness and skill necessitated the use of a categorical performance rating; otherwise global scores would have simultaneously served as outcome and predictor. I used the categorical performance score for both experience and skill-based analyses to increase consistency across analyses.

5.2 Results

To strengthen my claim that the surgeons are aware of their skill level, I demonstrate that confidence and performance are related, and that the confidence and performance share the same predictors. To demonstrate the relationship between events in the world to update one's self-perceived skill over time, I examined whether surgeons' self-confidence responded to likely indicators of skill level (e.g., training and simulated surgical performance). Further, I examine whether surgeons can monitor performance regardless of experience or skill level via the relationship between self-confidence and performance across experience and global scores.

The Maryland study gathered self-confidence ratings of the surgeons' anatomical knowledge and perceived ability to perform the procedure before and after every procedure[7]. In addition, the Maryland evaluators rated the surgeons' anatomical

[7] The available pairs for correlations and/or t tests were limited due to missing self-confidence ratings for some participants, resulting in low df for some tests. All available data was used for the analyses involving self-confidence ratings unless otherwise specified.

knowledge along with providing a global score. These data points served as my main predictor and outcome variables as I examined whether pre-procedure confidence predicted actual performance, as well as whether confidence changed in response to the success/failure of the procedure.

5.2.1 The relationship between a priori confidence and performance.

I sought to determine whether the surgeons demonstrated any awareness of their skill level, indicated by a relationship of self-confidence ratings of knowledge and ability with performance scores. I first correlated surgeons' self-confidence ratings with their performance scores in order to explore how well the surgeons were able to predict their own performance. I then identified shared predictors of confidence and outcome scores in order to rule out any spurious effects (e.g., the surgeons were confident in their skill for reasons unrelated to actual drivers of performance, leading to an artifactual link between confidence and performance).

5.2.1.1 Self-confidence is positively correlated with performance.

I evaluated the surgeons' self-awareness based on correspondence between self-confidence ratings and evaluator judgments of anatomical knowledge and performance. I correlated the surgeons' pre-procedure confidence in their anatomical knowledge of the shoulder/axillary region with the evaluators' overall rating of their understanding of anatomy in the axillary region. This correlation was significant ($r(38) = 0.64$, $p < 0.01$). I likewise correlated the surgeons' pre-procedure confidence in their ability to perform the procedure with the Maryland global score. This correlation was also significant ($r(38) = 0.52$, $p < 0.01$). These correlations indicate that the surgeons were at least broadly aware of both their relative levels of anatomical knowledge and overall skill.

5.2.1.2 Confidence and performance share predictors.

To bolster the claim that confidence reflected an awareness of skill, I investigated whether confidence and performance were influenced by the same predictors by correlating surgeons' pre-procedure confidence in both their anatomical knowledge and ability to perform the procedure with the individual Maryland and WSU predictors. These correlations are found in Table 5.1. The surgeons' confidence does indeed appear to be driven by many of the variables that predicted overall performance scores in Chapter 4 (Q8S1, Q8S2, and scores on the monitoring and strategy factors), implying at least some degree of awareness of the important elements of surgical skill and one's ability relative to those elements.

Table 5.1
Correlations (with df) between pre-procedure confidence ratings and the individual Maryland and WSU predictors.

Variable	Pre-procedure confidence in anatomical knowledge	Pre-procedure confidence in ability to perform the procedure
Goal Establishment		
Q1 (suspected injuries)	0.02 (38)	0.05 (38)
Q3 (additional studies)	0.23 (38)	0.19 (38)
Q12 (pitfalls)	0.23 (38)	0.14 (38)
Monitoring factor	**-0.53 (38)**	**-0.41 (38)**
Goal Enactment		
Q7 (landmarks and incision)	0.24 (38)	0.07 (38)
Q8S1 (procedure steps)	**0.61 (38)**	**0.48 (38)**
Q8S2 (technique points)	**0.53 (38)**	**0.53 (38)**
Q9 (expert operative field maneuvers)	**0.55 (38)**	**0.55 (38)**
Instrument change factor	0.05 (30)	-0.04 (30)
Strategy factor	-0.29 (38)	**-0.33 (38)**
Deliberate behavior factor	-0.11 (38)	-0.03 (38)
Oddities factor	0.01 (33)	-0.07 (33)

Note: Correlations in bold are significant at $p < 0.05$.

5.2.2 Confidence changed in response to information from the world.

In this section I examine how the surgeons' self-confidence changed in response to events in the world informative of one's skill level. In particular, I examined whether confidence responded to training (and thus presumably increases in skill level). I also examined whether confidence changed in response to actual rated performance, both of the surgeons' anatomical knowledge and also of their ability to perform the procedure.

5.2.2.1 Confidence increased in response to training.

I examined confidence changes in response to training via a series of paired samples t-tests to see if the residents' confidence in their anatomical knowledge and ability to perform the exposure changed in response to ASSET training (and thus presumably increases in skill level). Pre-procedure self-confidence ratings in anatomical knowledge improved after ASSET training ($t(12) = -3.61$, $p < 0.01$), as did pre-procedure confidence in ability to perform the procedure ($t(12) = -2.25$, $p = 0.04$). In order to determine whether this increase was due to improved skill or simply due to a psychological boost from the training, I tested whether performance (measured by Maryland global scores) improved after ASSET training. I found that performance after ASSET training was in fact improved relative to performance prior to training ($t(39) = -8.48$, $p < 0.01$). Together, these findings indicate that the surgeons' confidence increased in response to training (and by inference increased in response to changes in skill level).

5.2.2.2 Performance influenced confidence in ability, but not confidence in knowledge.

To more directly investigate the claim that confidence changed in response to skill, I also examined whether the surgeons' confidence was responsive to their actual performance, both in terms of scored levels of anatomical knowledge and in terms of performance on the procedure. I ran a regression using the evaluators' rating of the surgeons' overall anatomical understanding to predict the surgeons' post-procedure anatomical confidence, while controlling for the surgeons' pre-procedure confidence. The overall model was significant (**Model 5.1**; R^2 = 0.46, F(2,34) = 14.70, p < 0.01). While pre-procedure anatomical confidence was a significant predictor of post-procedure confidence (β = 0.58, t(35) = 3.381, p < 0.01), their rated understanding of the surgical anatomy was not (β = 0.14, t(35) = 0.80, p = 0.43). It therefore does not appear that the surgeons' confidence in their anatomical knowledge changed in response to their actual anatomical knowledge.

I next ran a similar regression to examine whether the surgeons' confidence in their ability to perform the procedure was responsive to performance. I used the Maryland global score to predict surgeons' post-procedure confidence in their ability to perform the procedure, controlling for their pre-procedure confidence ratings. The overall model was again significant (**Model 5.2**; R^2 = 0.47, F(2, 35) = 15.50, p < 0.01). As before, pre-procedure confidence significantly predicted post-procedure confidence (β = 0.43, t(35) = 2.84, p = 0.01). This time, however, the Maryland global score was also a significant predictor of post-procedure confidence ratings (β = 0.34, t(35) = 2.28, p = 0.03), indicating that the surgeons' post-procedure confidence in their ability to perform the procedure was responsive to how well they had actually performed during the procedure. Because the surgeons were not actually told their performance scores, this finding also provides further evidence of a monitoring function allowing the surgeons to be aware of their own performance.

5.2.3 The influence of experience and skill on the predictive value of self-confidence ratings.

I have established that the surgeons appear to be aware of their own skill level and that confidence responds to feedback from the world. To further enable surgeons to calibrate their level of self-monitoring, however, surgeons must be able to determine their skill level across different levels of experience and performance. Two analyses explore this question: the consistency of awareness across levels of experience and the consistency of awareness across levels of performance.

5.2.3.1 Awareness is consistent across levels of experience.

I explored whether more experienced surgeons were better judges of their own performance than less experienced surgeons, using a) ratings of anatomical knowledge and b) global performance scores as outcome measures. I used regression to test for an interaction between confidence ratings and the surgeons' career experience to see if the relationship between confidence ratings and outcome measures changed as a function of experience. A series of 4 models examines the questions of: a) whether anatomical self-confidence ratings and career experience predict evaluator ratings of anatomical knowledge, b) whether there is an interaction between anatomical self-confidence ratings and career experience in predicting evaluator ratings of anatomical knowledge, c) whether pre-procedure procedural self-confidence ratings and career experience predict Maryland global scores, and d) whether there is an interaction between pre-procedure procedural self-confidence ratings and career experience in predicting Maryland global scores.

5.2.3.1.1 Ratings of anatomical knowledge.

I generated a model predicting evaluators' ratings of the surgeons' anatomical knowledge using pre-procedure anatomical self-confidence ratings and the surgeons' career experience. This model was significant (**Model 5.3**; $R^2 = 0.41$, $F(2, 35) = 12.06$, $p < 0.01$). Pre-procedure anatomical confidence predicted evaluators' ratings of anatomical knowledge ($\beta = 0.59$, $t(35) = 4.41$, $p < 0.01$), but career experience did not ($\beta = 0.16$, $t(35) = 1.23$, $p = 0.23$). I then added the interaction term to the model, which did not increase the variance accounted for (**Model 5.4**; R^2 = 0.44, F(3,34) = 8.76, p < 0.01; F change (1,34) = 1.68, p = 0.20), and the interaction term was not a significant predictor (β = -0.22, t(34) = -1.30, p = 0.20), indicating that the relationship between self-confidence ratings of anatomical knowledge and evaluators' judgments of anatomical knowledge did not change with career experience (i.e., surgeons with more career experience were not more accurate in their assessments of their own anatomical knowledge).

5.2.3.1.2 Global performance scores.

I also tested for an interaction between procedural confidence ratings and the surgeons' years of experience to see if the relationship between confidence ratings and ratings of performance changed as a function of career experience. The model predicting Maryland global scores using pre-procedure procedural confidence ratings and the surgeons' experience was significant (**Model 5.5**; R^2 = 0.25, F(2, 35) = 5.73, p = 0.01). Pre-procedure procedural confidence predicted Maryland global scores (β = 0.50, t(35) = 3.23, p < 0.01), but career experience did not (β = 0.01, t(35) = 0.05, p = 0.96). The interaction term did not increase the variance accounted for (**Model 5.6**; R^2 = 0.29, F(3,34) = 4.59, p = 0.01; F change (1,34) = 1.99, p = 0.17), and the interaction term was not a significant predictor (β = -0.28, t(34) = -1.41, p = 0.17), indicating that experience did not affect the relationship between confidence and performance.

5.2.3.2 Awareness is consistent across levels of performance.

As with experience, I examined whether more skilled surgeons were better judges of their own performance than less skilled surgeons. I again tested for interactions between confidence ratings and global performance scores using a) ratings of anatomical knowledge and b) global performance scores as outcome measures. A series of 4 models examines the questions of: a) whether anatomical self-confidence ratings and relative performance predict evaluator ratings of anatomical knowledge, b) whether there is an interaction between anatomical self-confidence ratings and relative performance in predicting evaluator ratings of anatomical knowledge, c) whether pre-procedure procedural self-confidence ratings predict Maryland global scores, and d) whether there is an interaction between pre-procedure procedural self-confidence ratings and relative performance in predicting Maryland global scores.

5.2.3.2.1 Ratings of anatomical knowledge.

I also investigated whether better performers were more attuned to their anatomical knowledge. Because I did not want to use global scores to predict global scores, I first separated the surgeons into performance tiers: novice (more than 1 SD below the mean Maryland global score), journeyman (within 1 SD of the mean Maryland global score), and expert (more than 1 SD above the mean Maryland global score). I then predicted evaluators' ratings of the surgeons' overall understanding of axillary anatomy using pre-procedure anatomical confidence and performance tier in the first step, and the interaction term in the second step. The first model predicting ratings of the surgeons' overall understanding of axillary anatomy using pre-procedure anatomical confidence ratings was significant (**Model 5.7**; $R^2 = 0.70$, $F(2,37) = 42.72$, $p < 0.01$). Pre-procedure anatomical confidence predicted ratings of anatomical knowledge ($\beta = 0.25$, $t(37) = 2.26$, $p = 0.03$), as did performance tier ($\beta = -0.66$, $t(37) = -5.96$, $p < 0.01$). The interaction term did not increase the variance accounted for (**Model 5.8**; $R^2 = 0.71$, $F(3,36) = 29.54$, $p < 0.01$; F change $(1,36) = 1.66$, $p = 0.21$), and the interaction term was not a significant predictor ($\beta = 0.47$, $t(36) = 1.29$, $p = 0.21$), again indicating that the relationship between confidence and performance did not change across skill level.

5.2.3.2.2 Global performance scores.

I next tested for an interaction between the surgeons' self-reported confidence in their ability to perform the procedure and their performance. I constructed a model predicting the Maryland global scores using the surgeons' self-rated confidence in their ability to perform the exposure in the first step, then the interaction term between confidence and performance tier in the second step (performance tier was omitted from these models as it was inherently correlated with the global score). The model predicting Maryland global scores using pre-procedure procedural confidence ratings was significant (**Model 5.9**; $R^2 = 0.27$, $F(1, 38) = 14.30$, $p < 0.01$). Pre-procedure procedural confidence predicted Maryland global scores ($\beta = 0.52$, $t(38) = 3.78$, $p < 0.01$). The interaction term did not increase the variance accounted for (**Model 5.10**; $R^2 = 0.28$, $F(2,37) = 7.15$, $p < 0.01$; F change $(1,37) = 0.27$, $p = 0.60$), and the interaction term was not a significant predictor ($\beta = 0.30$, $t(37) = 0.52$, $p = 0.60$), indicating that the relationship between confidence and performance did not change as a function of the surgeons' skill level.

5.3 Discussion

My results indicated the surgeons appear aware of their performance generally (evidenced by a correlation between self-confidence and global scores). Surgeons demonstrated this awareness across levels of experience and levels of performance. Further, the surgeons are able to update their self-confidence based on events in the world (evidenced by changes in confidence in response to training and performance scores). The execution of skilled behavior in context demands some form of control in order to allow for the nuances of any particular situation. Automatically generated responses must be monitored for their fit with the situation and must be interruptible to allow for adjustment. Experts must adaptively utilize this interaction between explicit and implicit processes to their advantage when solving problems in context. The ability of the surgeons to judge their own skill facilitates this process by allowing them to allocate cognitive resources appropriately.

5.3.1 Some skills may not require career experience to develop.

One notable finding from the analyses presented above is that although both novice and experienced surgeons were aware of their performance, experience did not strengthen the relationship between confidence and performance (i.e., more experienced surgeons were not more accurate in their self-assessments). This finding strongly implies that monitoring behaviors are not learned over time, and raises the possibility that other skills may not require years of career experience to develop either. This issue is explored further in Chapter 6.

5.3.2 Sampling issues.

Several of my analyses suffer from low sample sizes and/or range restriction. These factors could easily have led to several of my null results due to low power, or could have led to spurious findings due to sampling error. My results are therefore best cast as suggestive. However, as discussed in section 2.2.1, the available sample is still quite impressive within the surgical domain. Despite the limitations of my analysis, they still serve to establish preliminary findings as a point of departure for future work.

5.3.3 Future research.

This chapter explored the cognitive penetrability of skilled performance via the surgeons' awareness of their skill level. Future research should address the contribution of self-awareness to monitoring in the execution of the procedures. I hypothesize that self-awareness allows the surgeons to calibrate their monitoring processes. Self-awareness is a component of cognitive penetrability, allowing the surgeon to devote appropriate resources to monitoring. Surgeons must monitor their procedures constantly (Dunphy & Williamson, 2004). Goals frame action selection (Huhn, Potts, & Rosenbaum, 2016). The integration of explicit monitoring with unconscious motor processes can lead to more efficient psychomotor behaviors (Shah, Barto, & Fagg, 2013).

However, monitoring is not without cost. Surgeons must be aware of their own performance in order to calibrate their self-monitoring processes. Devoting unnecessary effort to self-monitoring would likely slow the procedure and take cognitive resources away from other aspects of task execution. However, devoting too little effort to self-monitoring may lead to error. Because novice and expert surgeons alike are able to predict their own performance, confidence changed in response to the relative success of the procedure, and surgeons across performance levels demonstrated an awareness of skill, I believe it is plausible that the surgeons are indeed able to use conscious processes to adjust how closely they monitor their own behaviors during surgery. The negative relationship between the *monitoring* factor and performance observed in Chapter 4 is broadly consistent with this notion, although it must be kept in mind that the *monitoring* factor captures products of monitoring rather than monitoring itself – the better surgeons may still have been monitoring themselves but had fewer errors and less uncertainty to capture.

Chapter 6 – The Role of Experience in the Development of Expertise

6.1 Introduction

Surgical studies of expertise typically weight experience very heavily in their operationalizations of expertise. Studies explicitly define surgical experts as *experienced* surgeons with consistently better outcomes than other surgeons (Sadideen et al., 2013; Schaverien, 2010; e.g., Tien et al., 2015). However, the typical conceptualization both entrenches experience as a measure of competence and emphasizes outcomes. This is in contrast to Shanteau's notion of expertise, which focuses on relative performance and process. The medical perspective also ignores the possibility of high-performing inexperienced surgeons and poorly-performing experienced surgeons. By ignoring these possibilities, medicine is unable to describe fully the trajectory of skill development and likely misses behaviors that facilitate expert performance. Measures of skill must account for process as well as outcome in identifying expert performers. *How* a result is achieved can allow finer distinctions to be made among surgeons than outcome scores alone.

6.1.1 Experience as expertise.

Assuming that an individual is an expert purely due to certification or time on task is risky (Dunphy & Williamson, 2004). Novices can excel and even experienced practitioners may adopt flawed processes. The structure of knowledge is more important to performance than experience (Bradley, Paul, & Seeman, 2006). The nature of practice is more important than the absolute amount (Alderson, 2010; Ericsson, 2014).

However, an experience-based operationalization is not completely baseless. Experienced surgeons tend to complete procedures more quickly than junior and novice surgeons, and tend to demonstrate lower error rates and less variability in performance (Gallagher et al., 2001). This chapter explores the relative benefits of experience vs. training, and identifies areas that more experienced surgeons may differ from their less experienced counterparts despite similar outcomes.

6.1.2 The current analysis.

I sought to further investigate the current conceptualization of expertise by examining the contribution of experience to skilled performance. Below I use experience to account for Maryland global scores, examining the full sample of surgeons as well as comparing post-ASSET residents to attending surgeons. Limiting some analyses to the post-ASSET residents and attending surgeons serves to isolate the effect of experience, without the confound of the ASSET training intervention. I next explore the *necessity* of experience for expert-level performance by comparing residents and attending surgeons within the group of highest performers and after ASSET training. Finally, I examine the impact of training rather than experience, analyzing the contributions of prior training to performance and the benefits of the ASSET course itself. Throughout the analyses, I also identify differences among experience groups in outcome measures (the Maryland global score) and process measures (the individual Maryland and WSU predictors) to illustrate the relative ability of each to identify performance differences.

My analyses indicate that although experience does predict performance, the relationship is complex. The best residents were able to perform nearly identically attending surgeons, and post-ASSET residents as a group achieved similar Maryland global performance scores to the attending surgeons. Though the attending surgeons were still faster than even the best residents, experience does not appear to be totally necessary (much less sufficient) to achieve high levels of performance. Training (either before

ASSET or from the ASSET course itself) allowed residents to achieve levels of performance strikingly similar to their more experienced colleagues. Process-based measures were able to identify differences between post-ASSET residents and attending surgeons, while outcome-based measures could not.

6.2 Results

6.2.1 Accounting for global scores using experience.

As many investigations of surgical expertise have focused on experience as a key indicator of surgical skill, I sought to examine the contribution of experience to Maryland global scores. Because part of the Maryland parent study involved a pre-post training intervention in addition to a group of experienced experts, ASSET training status and experience are confounded (i.e., there is little to no difference in experience between the pre and post ASSET training groups, but the more experienced experts were counted as their own training phase). I therefore decided to conduct some analyses using only the post-ASSET training procedures and the expert procedures. I reasoned that analyzing the post-ASSET procedures in relation to the expert procedures would be more informative regarding whether experience alone provides any performance improvements beyond the ASSET course. The data set used (full or post-ASSET vs. attending surgeon) is identified in the description of each analysis.

6.2.1.1 Full data set.

6.2.1.1.1 Experience is positively associated with global scores overall.

I first correlated years of experience with global scores. Years of experience was positively associated with scores on the Maryland global outcome measure ($r(84)$[8] = 0.23, p = 0.04). This analysis provides a gross view of whether experience relates to

[8] Years of experience was missing for some participants, leading to fewer pairs and slightly lower df.

performance, but years of experience was largely unchanged between pre-ASSET and post-ASSET procedures among the residents. I next ran a series of regression models to determine how well experience accounted for variance in the Maryland global scores when controlling for ASSET training status. I generated regression models predicting Maryland global scores that entered training status (pre-ASSET, post-ASSET, or expert) in step 1 and surgeons' years of experience in step 2.

6.2.1.1.2 Experience is negatively associated with global scores when controlling for training status.

When predicting the Maryland global outcome scores, training phase by itself accounted for a significant proportion of the variance (**Model 6.1;** $R^2 = 0.44$, $F(1, 84) = 37.01$, $p < 0.01$). Training status was a significant predictor of Maryland global outcome scores ($\beta = 0.66$, $t(84) = 8.15$, $p < 0.01$). Surgeons' years of experience accounted for a significant additional proportion of the variance when controlling for training status (**Model 6.2;** $R^2 = 0.48$, $F(2, 83) = 37.91$, $p < 0.01$; F change $(1, 83) = 5.74$, $p = 0.02$). Both training status and years of experience predicted Maryland global outcome scores in model 6.2, though the results indicate that the most experienced surgeons actually had *lower* global outcome scores when controlling for training status (Table 6.1). This possibility is discussed in greater detail in the next section.

Table 6.1
Standardized beta weights for each of the variables in the final model using experience to predict Maryland global outcome scores.

Model	Variable	*β*	*t*	*p*
1	Training phase	0.66	8.15 (88)	< 0.01
2	Training phase	0.80	8.22 (87)	< 0.01
	Years of experience	-0.23	-0.40 (87)	0.02

6.2.1.2 Post-ASSET residents vs. attending surgeons.

I further investigated the relationship between experience and the Maryland global score using regression in a manner similar to that described above, but using only the post-ASSET and expert procedure data.

6.2.1.2.1 Experience demonstrates a nonlinear relationship with performance after training.

I generated a regression model with years of experience predicting Maryland global scores. The model with only experience did not significantly predict global outcome scores (**Model 6.3;** $R^2 = 0.01$, $F\ (1, 45) = 0.23$, $p = 0.64$) after ASSET training. However, I noted apparent nonlinearity in the residual plot (as well as a plot of surgeons' experience x Maryland global score; Figure 6.1), prompting me to investigate the possibility of a nonlinear trend in the data. The quadratic model for the data was not significant, (**Model 6.4;** $R^2 = 0.10$, $F\ (2, 44) = 2.43$, $p = 0.10$). Despite this, years in practice was a significant negative predictor of Maryland global scores in model 6.4 ($\beta =$ -0.93, $t(44) = -2.15$, $p = 0.04$).

Closer inspection of Figure 6.1 indicated that one of the attending surgeons was a possible outlier compared to the other attending surgeons, as he was the only experienced surgeon to have a negative z score for performance. This low data point may have pulled the curve down and created a false trend in the data set. I therefore re-ran the analysis, excluding this surgeon. This time, the quadratic model for the data was significant, (**Model 6.5;** $R^2 = 0.13$, $F\ (2, 43) = 3.31$, $p = 0.05$). Years in practice was again a significant negative predictor of Maryland global scores in model 6.5 ($\beta = -0.89$, $t(43) = -$ 2.24, $p = 0.03$).

This observed negative relationship between years of experience and Maryland global scores may be an accurate description of an effect in the world, or it may be due to artifacts of the ASSET evaluation procedure. I discuss each possibility in turn.

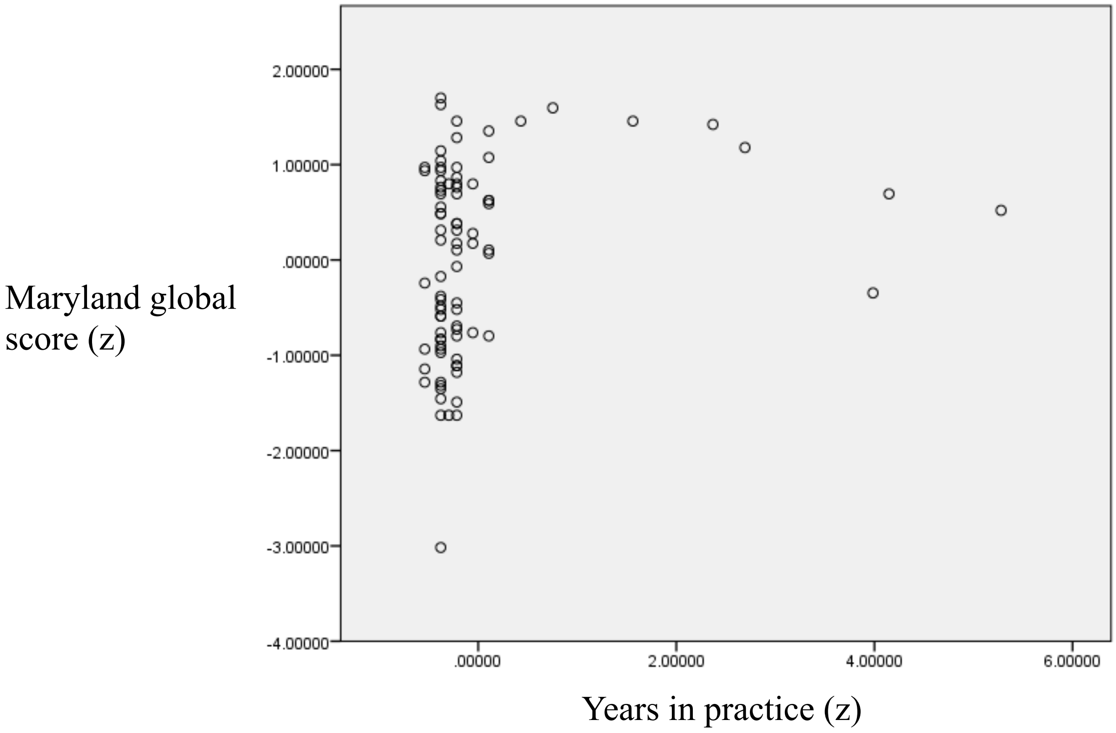


Figure 6.1. Surgeons' years of practice and associated Maryland global scores.

6.2.1.2.2 Hypothesized explanations for the nonlinear relationship between experience and performance.

First, the observed trend could be "real". If so, it seems plausible to think that the surgeons in the middle range of experience may have a good balance of knowledge and the ability to implement that knowledge in the world. Conversely, the surgeons with less experience may know *what* to do, but have not necessarily mastered *how* to do it (at least to the level of their more experienced counterparts). Another possibility is that surgeons at the high end of experience may start to perform more poorly due to too great a training retention interval, such that their knowledge or skills are no longer current given the rare nature of the studied procedures. Older physicians (correlated with experience) tend to satisfice more and think less analytically as well, which can impair performance on some tasks (Djulbegovich et al., 2014).

Alternatively, the nonlinear trend may have been due to more experienced evaluators. Due to the nature of the study, the evaluators had more experience in judging performance during the post-ASSET procedures than during the pre-ASSET procedures. In contrast, because the expert attending surgeons were evaluated before any of the residents, the evaluators had the least evaluation experience for these procedures. Thus, when analyzing the post-ASSET procedures and the attending surgeons' procedures, I examined the procedures for which the evaluators had the most and least evaluation experience, respectively. The observed pattern in scores may be due to evaluators' own experience differences rather than the surgeons'. This is addressed further in section 6.3.1.

The results of this analysis indicate that although experience predicts performance, the relationship is not straightforward. More experience does not necessarily indicate better performance. In order to better clarify the benefits conferred by experience, I examined differences between the highest-performing residents and the attending surgeons.

6.2.2 Investigating the necessity of experience for expert performance.

In order to identify any benefits gained via experience rather than training, I examined whether qualitative differences may exist between high-performing residents (regardless of ASSET status) and the more experienced attending surgeons (i.e., process may vary based on experience even though outcome may not). I separated the surgeons into novice (more than 1 SD below the mean Maryland global score), journeyman (within 1 SD of the mean Maryland global score), and expert (more than 1 SD above the mean Maryland global score) tiers based on the Maryland global scores and looked at differences between resident and attending surgeons within the top tier. The expert group contained eight resident surgeons and five attending surgeons.

6.2.2.1 Lack of process differences among the top performers.

I used a series of independent samples t-tests to compare the residents and the attending surgeons in the expert group across the process-oriented Maryland and WSU variables[9]. None of the predictors demonstrated significant differences between selected residents and the attending surgeons (Table 6.2). Although experience did not appear to affect surgical process per se, it did facilitate faster procedures among the attending surgeons. Overall, it appears that experience by itself does not confer any additional skill that cannot be acquired through training, as indicated by the predictors used here. To identify the benefits offered by training, I examined how ASSET training appeared to improve some residents' performance to the point of being comparable to attending surgeons.

[9] The journeyman and novice performance tiers did not contain a sufficient number of attending surgeons to serve as a basis of comparison to the residents.

Table 6.2

t-test results comparing resident and attending surgeons for WSU and Maryland predictors within the "expert" performance tier.

Variable	Mean Diff. (Res - Att)	*t(df)*	*p*
Goal Establishment			
Q1 (suspected injury)	0.58	0.81 (12.00)	0.44
Q3 (additional studies)	-0.71	-1.43 (12.00)	0.18
Q12 (pitfalls)	-0.57	-0.83 (12.00)	0.39
Monitoring factor	-0.11	-1.27 (12.00)	0.23
Goal Enactment			
Q7 (landmarks and incision)*	-0.19	-0.71 (11.05)	0.49
Q8S1 (steps of the procedure)	-0.18	-0.96 (12.00)	0.36
Q8S2 (technique)*	-0.30	-2.00 (10.96)	0.07
Q9 (expert operative field maneuvers)*	-0.97	-2.31 (07.00)	0.06
Instrument change factor	0.41	1.76 (11.00)	0.11
Strategy factor	0.31	1.41 (12.00)	0.18
Deliberate behavior factor*	-0.48	-1.14 (05.29)	0.31
Oddities factor	0.12	0.38 (11.00)	0.71
Total completion time	0.34	2.51 (12.00)	**0.03**

Note: *Some variables failed Levene's test for equality of variances, leading to altered df.

6.2.2.2 Process differences between post-ASSET residents and attending surgeons.

In order to better examine whether ASSET training truly allows residents to perform as well as attending surgeons, I ran a series of independent samples t-tests between all post-ASSET residents and the group of attending surgeons. Although global scores as a whole (i.e., outcome measures) were not significantly different between post-ASSET residents and attending surgeons ($t(48) = -1.27$, $p = 0.21$), results indicated several differences on individual predictors (i.e., process-based measures). Attending surgeons demonstrated better scores on the Maryland variable Q8S2 (technique points) and Q9 (expert operative field maneuvers). Attending physicians demonstrated lower scores on the WSU *instrument change* factor and the *strategy* factor (Table 6.3). Among these variables, the Maryland Q8S2 variable and the WSU *instrument change* and *strategy* factors were significant predictors of performance in chapter 4. Although the attending surgeons did not demonstrate any process-related differences from the resident surgeons within the best-performing group (section 6.2.2.1 above), it appears that the attending surgeons do display qualitatively different processes than the less experienced post-ASSET residents as a whole, despite similar overall outcome scores.

Table 6.3
t-test results comparing post-ASSET resident and attending surgeons for WSU and Maryland predictors.

Variable	Mean Diff. (Res - Att)	*t(df)*	*p*
Goal Establishment			
Q1 (suspected injury)	-0.23	-0.60 (48.00)	0.55
Q3 (additional studies)	-0.56	-1.96 (48.00)	0.06
Q12 (pitfalls)*	0.03	0.06 (10.35)	0.95
Monitoring factor	0.00	0.03 (48.00)	0.98
Goal Enactment			
Q7 (landmarks and incision)	-0.01	-0.04 (48.00)	0.97
Q8S1 (steps of the procedure)	-0.12	-0.60 (48.00)	0.55
Q8S2 (technique)*	-0.93	-4.87 (25.59)	**< 0.01**
Q9 (expert operative field maneuvers)*	-1.04	-4.90 (36.48)	**< 0.01**
Instrument change factor	0.49	2.86 (47.00)	**0.01**
Strategy factor	0.44	2.06 (48.00)	**0.04**
Deliberate behavior factor*	-0.78	-2.10 (10.08)	0.06
Oddities factor	-0.50	-1.79 (47.00)	0.08

Note: *Some variables failed Levene's test for equality of variances, leading to altered df.

6.2.3 Examining the impact of training.

In addition to identifying pure time on task effects, I sought to determine whether the *nature* of a surgeon's experience affected performance scores, including prior training as well as ASSET training.

6.2.3.1 Prior training benefits performance.

I used the demographic information collected from the Maryland study (noted at the beginning of Chapter 3; prior cadaver-based training (yes or no), total hours spent in the cadaver lab since medical school, and total hours in the open skills lab since medical school) as predictors to determine whether specific types of experience prior to ASSET training were able to predict performance. I used a similar approach as above, generating regression models predicting the Maryland global outcome measure using training status (pre-ASSET, post-ASSET, or attending physician) in the first step and training or experience prior to ASSET training in the second step.

When predicting Maryland global scores, training phase by itself accounted for a significant proportion of the variance (**Model 6.6**; $R^2 = 0.43$, $F(1, 85) = 63.32$, $p < 0.01$). Training status was a significant predictor of Maryland global scores ($\beta = 0.65$, $t(85) = 7.96$, $p < 0.01$). I next added whether the surgeons had taken cadaver based courses prior to ASSET and the number of hours the surgeon had spent in the cadaver or open skills labs since medical school to the model. The resulting model was significant but did not predict significant additional variance in Maryland global scores compared to model 1 (**Model 6.7**; $R^2 = 0.47$, $F(4, 82) = 18.20$, $p < 0.01$; F change $(3, 82) = 2.23$, $p = 0.09$). Despite the lack of significant additional variance accounted for by the additional variables overall, whether the surgeon had taken other cadaver-based courses before ASSET training did significantly predict Maryland global scores controlling for training status ($\beta = 0.19$, $t(82) = 2.31$, $p = 0.02$). The number of hours spent in the cadaver and open skills lab since medical school were not significant predictors (Table 6.4). Taking cadaver-based courses prior to the ASSET course appears to be beneficial to performance in the context of the Maryland evaluation, but simply spending time in surgical labs is not.

Table 6.4

Predicting Maryland global scores using training status and other types of experience.

Model	Variable	*β*	*t*	*p*
1	Training phase	0.65	7.96 (88)	< 0.01
2	Training phase	0.64	7.13 (85)	< 0.01
	Prior cadaver-based courses	0.19	2.31 (85)	0.02
	Hours in the cadaver lab	-0.01	-0.02 (85)	0.99
	Hours in the open skills lab	-0.09	-0.32 (85)	0.75

6.2.3.2 ASSET training benefits many aspects of performance.

Because ASSET training appears to allow at least some residents to perform on a level comparable to some attending surgeons (evidenced by similarity between residents and attending surgeons within the expert performance tier; section 6.2.2.1 above), I sought to explore the mechanism for this improvement by determining which individual skills benefited from ASSET training. I ran a series of dependent-samples t-tests on the Maryland and WSU predictors comparing the residents' pre-ASSET scores to the residents' post-ASSET scores. All of the process-oriented Maryland variables with the exception of Q1 (suspected injury) changed after ASSET training. Among the process-oriented WSU variables, scores on the monitoring factor were significantly lower after ASSET training than before ASSET training (Table 6.5). ASSET training appears to impact many of the process-oriented predictors in this study, though not enough to yield results comparable to the attending surgeons.

Table 6.5
t-test results comparing residents' pre- and post-ASSET training for WSU and Maryland predictors.

Variable	Mean Diff. (Pre - Post)	*t(df)*	*p*
Goal Establishment			
Q1 (suspected injury)	-0.23	-1.14 (39)	0.26
Q3 (additional studies)	0.42	2.08 (39)	**0.05**
Q12 (pitfalls)	-0.53	-3.13 (39)	**< 0.01**
Monitoring factor	0.78	4.48 (39)	**< 0.01**
Goal Enactment			
Q7 (landmarks and incision)	-1.24	-8.14 (39)	**< 0.01**
Q8S1 (steps of the procedure)	-1.39	-8.42 (39)	**< 0.01**
Q8S2 (technique)	-0.98	-6.81 (39)	**< 0.01**
Q9 (expert operative field maneuvers)	-0.53	-2.62 (38)	**0.01**
Instrument change factor	0.30	1.20 (19)	0.25
Strategy factor	0.25	1.80 (39)	0.08
Deliberate behavior factor	0.16	1.09 (39)	0.28
Oddities factor	0.39	1.82 (24)	0.08

6.3 Discussion

Experience and deliberate practice (Ericsson, Krampe, & Tesch-Romer, 1993) are believed to be the key requirements for expert-level skill to develop. Many domains, surgery included, traditionally operationalize expertise based at least in part on experience. The results of this study challenge the strong role of experience in the medical community's common view of expertise, although the possibility of sampling error (due to small sample sizes of attending physicians overall and surgeons within the expert performance tier) precludes concrete conclusions. Though not as pronounced as the limits in Chapter 5, many of my analyses in this chapter utilized truncated samples within the data set. This reduced power and potentially led to range restriction. Again, these findings are best cast as a preliminary point of departure for future work.

6.3.1 The relationship between experience and performance is complex.

Although my findings certainly indicate that experience contributes to performance (evidenced by the positive correlations between experience and global scores, as well as speed advantages for attending surgeons over even high-performing residents), the relationship is not as clearly defined as the medical literature often assumes. Several effects illustrate this point: A nonlinear relationship between experience and performance, overlapping resident-attending global performance scores after ASSET training, the general benefit of cadaver-based training on performance regardless of ASSET status, and the qualitative similarity between residents and attending surgeons within the best performance tier. Differences in process between experience cohorts persist, but are essentially absent among the top performers.

As mentioned in section 6.2.1.2.2, I believe that one possible explanation for the nonlinear relationship between experience and performance among the post-ASSET residents and attending surgeons is a confound between testing group and evaluator experience in conducting evaluations. The least experienced evaluators evaluated the attending surgeons, while the most experienced evaluators evaluated the post-ASSET residents. This issue and other potential effects of evaluator cognition on performance scores are addressed in Chapter 7.

6.3.2 Experience is not necessary for expert-level performance.

Some residents were able to achieve performance nearly indistinguishable from more experienced attending surgeons, and I noted multiple cases of poor performance within the attending physicians. Among the best performers, the only difference between attending surgeons and residents was time to completion. Overall it appears that the main benefit of experience is in the smoothness/speed of physical execution. All other measures (including the more general, process-oriented WSU predictors) in this study indicated that at least some residents can equal attending surgeons with the proper training. My data indicate that time on task may not be necessary for expert-level performance (and almost certainly is not sufficient).

Post-ASSET improvement notwithstanding, it seems that some behaviors still benefit from experience. The attending physicians demonstrated better scores on the Maryland variables of technique and operative field maneuvers than post-ASSET residents as a group and fewer *instrument changes* and *strategy* behaviors compared to these same residents. Nevertheless, differences between residents and attending surgeons within the top performance tier were more quantitative than qualitative. The best-performing residents (in the top performance tier) were not significantly different from the more experienced attending surgeons on any Maryland or WSU measures despite their relative inexperience. However, despite the dearth of technical or procedural differences between the best performing residents and the more experienced attending surgeons, the residents were still slower to complete the procedure. Experience certainly affects performance, but it appears that less-experienced surgeons are able to perform at near-expert level with the proper training.

6.3.3 Training benefits multiple aspects of expert performance.

As noted in the results section, cadaver-based training prior to the ASSET course was a significant predictor of performance, even controlling for ASSET training status, and almost all the Maryland behaviors (and some WSU factors) were impacted by the ASSET training course. ASSET training allowed residents to achieve global performance scores comparable to those of more experienced attending surgeons, despite some differences in scores on individual components of the evaluation form. Given that the axillary exposure portion of the ASSET course is relatively brief, such improvement in skills is impressive. I found that the WSU *monitoring* factor was also reduced by ASSET training. Changes to the *monitoring* factor after training are in line with the mediation described previously and provide converging evidence for the importance of these variables in characterizing expertise.

6.3.4 Global scores offer only a partial glimpse of performance.

My results indicate that process-based measures can be more useful in identifying expert behavior than outcome-based measures (or at least more helpful in distinguishing between groups of surgeons). Though outcome-based measures of performance (the Maryland global score) did not detect a difference between post-ASSET residents and attending surgeons, several differences between these cohorts were detected among the individual Maryland and WSU predictors. While global outcome scores remain a useful measure of gross performance, more fine-grained distinctions require the use of process-based measures.

Chapter 7 – Evaluator Cognition

7.1 Introduction

Ideally, most variance in performance scores would be attributable to trainee behavior, and variance attributed to the raters would be considered error (McGill, van der Vleuten, & Clarke, 2011). However, the trainee may only contribute 25% of the variance in rating scores (McGill, van der Vleuten, & Clarke, 2011). Interrater reliability is one of the prerequisites for a useful measurement tool. I established in Chapter 3 that Maryland global scores and several individual Maryland predictors were reliable when considered across the data set as a whole, despite the rejection of many Maryland and WSU items as unreliable. However, this metric may be too general to capture nuance within specific aspects of a data set. The context in which an evaluation is conducted, the assessment tool used, and characteristics of the evaluator can all influence performance measurement (Mitchell et al., 2014). Such influences on performance scores beyond the trainee must be accounted for in order to better isolate and assess expert performance. This chapter examines several influences on evaluator judgments in an effort to reduce their contribution to performance scores.

7.1.1 Evaluation context can impact performance ratings.

Contextual variables such as the body habitus of the cadaver or the method of evaluation (paper and pencil, electronic, etc.) can potentially affect scores. Exceptionally thin or obese patients may be more difficult to operate on, and evaluators must take this into account when judging performance. Likewise, evaluations using electronic tablets may be easier to navigate, may facilitate use of prior items as anchors for later items, or may make it more difficult to leave evaluation items blank.

7.1.2 The structure of the evaluation form may impact performance ratings.

Individual items may be difficult to judge independently due to prior items serving as anchors or points of reference. To the extent that items influence one another, the structure of the rating form itself and the order of evaluation items may impact scoring independently of the actual skill of the person being evaluated. If present, the influence of items on one another may be largely unavoidable. However, the evaluation process will be strengthened by an awareness of this influence and by efforts to mitigate its impact on scoring.

7.1.3 Evaluator cognition affects scoring in addition to trainee skill.

We like to believe that evaluators can measure performance objectively, with scores only influenced by the skill of the person being evaluated. However, generating a single global performance score requires evaluators to integrate multiple evaluation items according to idiosyncratic weighting criteria. An evaluator's cognitive processes inevitably impact scores. An evaluator's experience, background, or familiarity with the evaluation process may affect how they interpret and judge evaluation items and arrive at global scores. The criteria evaluators use to make their assessments and the consistency with which these criteria are applied may vary across evaluators as well.

7.1.4 The current analysis.

A design such as that employed in the Maryland evaluation study makes these issues even more relevant. The Maryland study utilized multiple evaluators, with varying backgrounds and experience that may lead to different scoring criteria between raters. Evaluations were not evenly divided among the evaluators, potentially giving some individuals greater influence over the overall data set than others. Further, reliability of individual evaluators is difficult to assess due to the lack of repeated measurements from evaluators rating the same procedure multiple times.

I have relied on the evaluators' subjective global ratings as a gold standard for my previous analyses because the holistic global scores provide the best opportunity to account for behaviors not necessarily included in the rating form. Such subjective ratings have proven to be reliable overall and correlated with other measures of performance (e.g., the WSU objective rank score). However, this subjectivity also leaves increased room for outside influences to affect the evaluators' judgments. I examined evaluator reliability, along with three sources of influence that may have impacted evaluators' judgment when evaluating the surgeons: influences stemming from the context of the evaluation, influences related to the specific evaluation tool itself, and influences internal to the evaluators themselves. This analysis served to further identify variables that should be accounted for to ensure that performance measures best capture skilled behavior. I utilized a variety of analysis techniques in order to best describe these influences.

I first evaluate interrater reliability within specific sections of the data set. This analysis provided a more detailed picture of how well evaluators agreed for specific parts of the Maryland study, rather than in the study as a whole. I next examine scoring influences related to context including the body type of the cadaver (thin, average, or obese), the method of evaluation (paper and pencil or tablet), whether the evaluation was based on only the cadaver or a combination of the cadaver and the surgical model, and the specific evaluator who provided the ratings. I adopted a Bayesian approach in order to account for the uneven distribution of evaluations across evaluators. I then examine the effect of item order on scoring, using time series analysis to determine the effect of prior items presented in the evaluation form on ratings of later items. Next, I examined the influence of individual-level variables among both the surgeons and evaluators. I adopted a multilevel approach to account for the nested nature of these data (surgeons within evaluators). Finally, I examined individual raters' scoring consistency using Cochran-Weiss-Shanteau (CWS) analyses. Together, this diverse array of analyses serves to identify issues of between-rater reliability, the effects of contextual influences on scoring, the effects of the evaluation tool on scoring, and the impact of the evaluators' own cognitive processes on scores.

7.2 Results

Measures of overall reliability do not account for variation within specific segments of the data set, such as different skill levels, or changes in the evaluators over time. I examined Krippendorff's alpha across different segments of the data set, examining reliability across levels of trainee performance and evaluator experience. My results indicated that evaluators were not equally consistent at all times across all skill levels. In an effort to examine MADM (how evaluators combined multiple pieces of information to arrive at global ratings), I examined how a) the external context of the evaluation, b) the structure of the rating form, and c) variation in judgment across evaluators (including across demographic factors and trainee skill) may have influenced how the raters evaluated the surgeons. Contextual factors and individual evaluator characteristics both appear to contribute to evaluator judgment when generating global rating scores.

7.2.1 Krippendorff's alpha indicates that reliability varies with trainee performance and evaluator experience.

Though Maryland global scores were reliable when the sample was considered as a whole, I examined whether evaluator consistency varied across levels of performance. To better isolate different levels of performance, I computed Krippendorff's alphas for the Maryland global scores within each of the performance tiers described in Chapter 6 (i.e., novices more than 1 SD below the mean, journeymen within 1 SD of the mean, and experts greater than 1 SD above the mean Maryland global score). Krippendorff's alphas for tiers 1, 2, and 3 were -0.42, 0.57, and -0.18, respectively. These negative alpha values indicate that although the evaluators agreed on global scores for the data set as a whole (alpha was 0.80), they did not agree at the more extreme performance ranges. In fact, such negative values indicate that they systematically *disagreed* (DeSwert, 2012). People seem to agree on middle-of-the-road levels of performance, but may have different criteria for what constitutes very good or very bad performance. Evaluators' judgments varied based on trainee performance.

Because performance was somewhat confounded with the evaluators' experience in performing evaluations (the experts were evaluated first, followed by pre-ASSET procedures, then post-ASSET procedures), I also examined Krippendorff's alpha for the Maryland global scores for each of these three phases of the study. Krippendorff's alpha for the first round of evaluations (the experts) was 0.49. The second round of evaluations (the pre-ASSET procedures demonstrated an alpha of 0.67. Finally, alpha for the post-ASSET set of procedures was 0.77. These increasing values suggest that the evaluators as a group became more consistent in their ratings over time. Krippendorff's alphas were likely lower among the high performers because the best scorers (the attending surgeons and the post-ASSET residents) were evaluated when the evaluators had the least and most experience. However, the improved consistency over time was not enough to make up for the inconsistency in the early evaluations.

7.2.2 External context of the evaluation affects evaluators' judgments.

I used two Bayesian models, both evaluating whether global scores were impacted by contextual factors other than the trainees' behaviors and skill (the rating form, body type and the use of a synthetic surgical model)[10]. The first model examined these variables in the context of the Maryland and WSU predictors, whereas the second model excluded the Maryland and WSU predictors. I conducted my analyses in JASP Version 0.8 Beta 5.

7.2.2.1 Model 1: examining contextual factors in the context of Maryland and WSU predictors.

I generated a Bayesian model using only the aggregated Maryland evaluation items (Q1, Q3, Q7, Q8S1, Q8S2, Q9, & Q12), retaining the items included in the strongest model. I next repeated this process using only the five WSU factor scores. The Maryland items Q8S1 (steps of the procedure) and Q8S2 (technique points), along with scores on the Instrument Change, Strategy, and Monitoring factors emerged from this analysis. These are the same variables that emerged as significant predictors of global scores in Chapter 4 (sections 4.2.1.1 and 4.2.1.2). Finally, I constructed a Bayesian model using the five predictors just described, along with the body type of the cadaver (thin, average, or obese), the method of evaluation (paper and pencil or tablet), whether the

[10] As this analysis was intended to examine evaluator cognition, only the subjective Maryland global outcome scores were analyzed in this analysis. The other outcome measures (Maryland IPS scores and WSU objective rank) were not assessed and are not included in the appendices.

evaluation was based on only the cadaver or a combination of the cadaver and the surgical model, and the specific evaluator who provided the ratings.

Results indicated the strongest evidence in favor of the Bayesian model retaining only the specific Maryland items Q8S1 (steps of the procedure) and Q8S2 (technique points), with a Bayes Factor (BF_{10}) of 1.447 $x10^{27}$. Bayes factors are relative values indicating the level of support for one model over another. Models with larger Bayes Factors have more support than models with lower Bayes Factors. The absence of WSU variables is not surprising given the mediation established in Chapter 4. Other Bayesian models with strong evidence included the same Maryland evaluation items plus whether the evaluation was based on only the cadaver or on a combination of the cadaver and a surgical model (BF_{10} = 1.271 $x10^{27}$), and the two Maryland items plus the evaluation method used (paper and pencil or tablet; BF_{10} = 1.066 $x10^{27}$).

7.2.2.2 Model 2: examining contextual factors alone.

In order to better determine the effects of external variables in the absence of evaluation items, I also examined a Bayesian model with only the external predictors described above. Evaluation items were excluded from this model in order to avoid the impact of potential correlations among predictors (WSU and Maryland predictors were correlated, and correlations between external predictors and individual evaluations items are plausible if the external influences affected scoring on individual items). Similar to the above analyses, the Bayesian model with the evaluation method used and whether the evaluation relied solely on the cadaver proved to be the strongest (BF_{10} = 16299.96). T-tests revealed that global scores for evaluations conducted with the tablet (later in the study) were higher than the evaluations conducted using paper-and-pencil ($t(162) = 2.47$, $p = 0.02$).

Performing the procedure on a surgical model prior to the cadaver may have affected Maryland global scores by providing an opportunity to practice. Scores on the cadaver-based procedure were higher than scores on the surgical model-based procedure when the model-based procedure was performed first ($t(17) = 2.17$, $p = 0.05$). Trainee performance on the surgical model may also have biased the later evaluation of the cadaver-based procedure by serving as an anchor for scores or allowing the trainee to get credit for demonstrating relative improvement.

7.2.3 The structure of the rating form does not influence how the raters evaluated the surgeons.

I also examined whether the specific order of the items on the evaluation form influenced evaluations, separately from the medium in which the evaluation form was utilized. Specifically, I hypothesized that prior items on the evaluation form may have generated an impression of the trainee that affected scoring on later items. I tested this possibility using time series analysis. I examined the autocorrelations among items (up to 16 lags) for individual evaluators within each procedure. The data were somewhat equivocal. Only 23 of the evaluations demonstrated any correlations at all, with no clearly apparent pattern in the autocorrelations across these evaluations as a whole. Likewise, individual evaluators did not demonstrate consistent patterns of correlation across their evaluations. However, some individual series demonstrated apparent structure.

In order to gain a better sense of trends at the aggregate level, I constructed box and whisker plots for each of the lags (i.e., a box and whisker plot for all of the lag 1 correlations, all of the lag 2 correlations, etc.). Each of these boxplots with the possible exception of lag 2 (Figure 7.1) indicated that the distribution of correlations was roughly centered around 0.00, leading me to conclude that the order of items on the rating form was unlikely to have an effect on scoring. The remaining plots can be found in Appendix K.

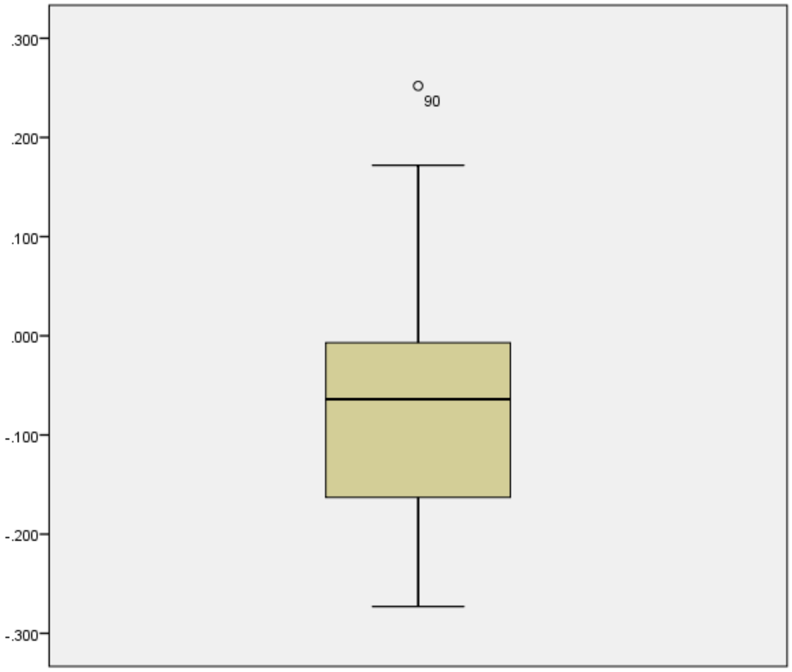


Figure 7.1. Lag 2 box plot for the time series analysis investigating the effect of the order of items on the evaluation form.

7.2.4 Variation in judgment across evaluators.

I examined the impact of influences related to the evaluators themselves. I examined demographic trends caused by sex effects, experience, and specialty. I also examined changes in evaluator consistency (both within and between evaluators) related to the skill of the surgeons being evaluated as well as the experience of the raters with the evaluation process. I did not find any demographic bias in the evaluators' judgments, but notable differences between evaluators emerged. Consistency in judgment appears to vary across evaluators depending on the skill of the surgeon under evaluation and the number of evaluations performed.

7.2.4.1 Demographic factors do not influence performance scores.

I sought to determine the impact of demographic variables that may have affected evaluators' judgments of performance. Because individual surgeons were nested within evaluators, I utilized a multilevel modeling approach. This approach accounts for shared variance due to similarities within higher-level groupings. At the first (surgeon) level, I examined variance attributable to the sex of the trainee and whether the trainee's handedness corresponded to the side of the cadaver from which the trainee operated. I suspected that it might have been easier or harder to perform a procedure based on whether the surgeon had to reach or adjust their positioning to use their preferred hand depending on their position relative to the cadaver and the simulated wound.

At the second (evaluator level), I examined the impact of evaluator demographics. Specifically, I examined the effect of evaluator sex, the evaluator's specialty (surgeon or non-surgeon), the evaluator's career experience, and the total number of evaluations performed by the evaluator.

Following the steps suggested by (Bliese, 2002), I first calculated an intraclass correlation coefficient (ICC) based on the null model predicting Maryland global scores. The ICC gives a measure of the amount of variance in Maryland global scores attributable to the evaluators. An ICC less than 0.10 indicates that a multilevel approach may not be necessary. The ICC based on the null model of Maryland global scores was 0.0001, indicating that only a tiny fraction of variance could be attributed to the evaluators. I therefore abandoned the multilevel approach in favor of linear regression to examine the impact of demographic artifacts on evaluators' judgments of global scores.

I predicted Maryland global scores using trainee-level variables (sex and handedness vs. side of the cadaver operated on) in step one of the model, and evaluator-level variables (sex, specialty, career experience, and total evaluations) in step two of the model. The first model predicting Maryland global scores using trainee-level variables was not significant (**Model 7.1**; R^2 = 0.00, F(2,154) = 0.18, p = 0.84). The individual trainee-level variables were not significant predictors of Maryland global scores. The model including evaluator-level variables did not account for additional variance in Maryland global scores (**Model 7.2**; R^2 = 0.02, F(6,150) = 0.39, p = 0.88; F change (4,150) = 0.50, p = 0.74), and the individual evaluator-level variables were not significant predictors.

7.2.4.2 Evaluator consistency varied with trainee skill.

I examined whether the evaluators differed in how consistently they evaluated trainees. In other words, whether some evaluators were able to more consistently differentiate among levels of performance and apply criteria consistently across trainees. I utilized the Cochran-Weiss-Shanteau (CWS) index in order to investigate this question. The CWS index evaluates evaluator expertise based on how well they discriminate between various stimuli, and how consistently they are able to make those distinctions (Weiss & Shanteau, 2014a). CWS scores are rater-specific, in contrast to Krippendorff's alpha, which provides a measure of between-rater reliability[11]. CWS scores can therefore provide insight into individual evaluators and converging evidence regarding the differences between evaluations at different levels of performance.

The CWS measure represents a relativistic view of expertise – it does not make assumptions about the accuracy of a given set of judgments, rather it is an index of the process used by the expert. In the absence of ground truth, however, such a process-based

[11] This also explains apparent contradictions between CWS results and Krippendorff's alpha results, particularly when examining the high-performing surgeons.

measurement is likely to provide a good description of judgment quality (Weiss & Shanteau, 2014a).

CWS is calculated as discrimination divided by inconsistency. Discrimination in this equation refers to a given person's ability to differentiate between the stimuli of a presented set, calculated as the variance among averaged responses to separate stimuli. Inconsistency refers to the person's ability to judge the same stimulus over multiple instances, calculated as the variance in responses to the same stimulus averaged across the set of stimuli (Weiss & Shanteau). Higher CWS scores therefore indicate better evaluator judgment, as one desires high variance between judgments of separate stimuli and low variance between judgments of the same stimulus.

To approximate repeated stimuli, I treated similarly-rated surgeons as equivalent stimulus presentations from the evaluators' perspective. This necessitated identifying the surgeons evaluated by each rater and dividing those surgeons into roughly equivalent performance categories for evaluation. I first identified the number of evaluations performed by each evaluator. Evaluators with less than 10 evaluations were excluded in order to ensure a minimum number of evaluations for the analysis. Four evaluators performed at least 10 evaluations. Because I needed to examine equivalent cases for each evaluator, I next separated the ASSET trainees into bins based on their Maryland global performance score. Those scoring less than 60 were placed in a "low performing" bin, and those scoring above 85 were placed in a "high performing" bin for analysis. The surgeons within each bin were treated as equivalent cases for the purpose of computing CWS index scores for each evaluator. One additional evaluator was eliminated due to having insufficient surgeons fall into both the "low performing" and "high performing" bins for analysis, leaving three evaluators with a sufficient sample to calculate CWS scores.

I treated each set of aggregated items as a stimulus (e.g., the aggregated subitems within question 1 counted as a single stimulus). The individual surgeons within the performance bins were treated as repetitions of those stimuli. Two CWS scores were calculated for each evaluator (one per performance bin). Table 7.1 lists these evaluators and their CWS scores within each bin.

Table 7.1
CWS scores for each evaluator for both low and high performing surgeons.

Evaluator	Low performers	High performers
8	64.05	238.54
10	43.80	285.01
15	189.70	370.37

Note: Higher scores indicate more consistent judgments.

A statistical test to compare F ratios (applicable to CWS scores) is available. However, such a test requires that the same set of stimuli be used for each evaluator (Weiss & Shanteau). Because our evaluators all evaluated a different set of surgeons, their CWS scores are not subject to statistical comparison. However, I note two distinct trends in the data. First, Evaluator 15 consistently had higher CWS scores than the other two evaluators, suggesting that some evaluators in the ASSET study may have been better than others at judging performance. Second, CWS scores were higher for the "high performing" bin than the "low performing" bin in all cases, suggesting that evaluators may have been better at judging good performance than poor performance.

7.3 Discussion

We like to think of evaluators as impartial observers, grading performance objectively based solely on the merits of the observed behavior. However, human evaluators are not dispassionate sensors capable of repeating the same outcome over and over again. Evaluators are subject to the same cognitive influences as anyone else and I found indications of extraneous influences on scoring that form the basis of recommendations for future performance evaluation. While demographic analysis provided encouragement regarding the ability of the evaluators to judge performance, reliability indices such as CWS scores and Krippendorff's alpha and analyses of contextual factors indicate that multiple factors beyond surgeon performance influence scores. Outside influences on scores notwithstanding, however, only a human with the requisite background knowledge can properly identify many of the behaviors identified in previous chapters as important to expertise. We must therefore strive to be aware of some of the external factors that can artificially impact outcome scores and do what we can to combat these extraneous influences.

7.3.1 Suggestions to mitigate context-specific influences on rater judgment.

Looking beyond the surgeons and evaluators themselves, I found evidence that contextual features affected scoring independently of the surgeons' performance. Most of the contextual factors that influenced scoring in my analyses were a product of the Maryland study design rather than factors likely to be found in real-world evaluations. Nonetheless, they have implications for more generalizable situations.

I did not have data available to examine whether prior evaluations of other trainees impacted later evaluations of different trainees within the same evaluator, but the model vs. cadaver data provide some level of insight into whether the evaluators were able to judge individual procedures in isolation. Bayesian modeling indicated that prior evaluation of performance on a surgical model may have influenced later cadaver-based evaluations. Although this situation is relatively artificial and likely unique to the

Maryland study, the results of my analysis suggest that the evaluators did not judge procedures independently of one another. I therefore recommend that a given evaluators' evaluations be spaced in time, or that different tasks be evaluated back-to-back to the extent possible in order to minimize any carryover effects between evaluations.

Bayesian analyses also indicated that the evaluation method (paper-and-pencil vs. tablet) affected scoring. The observed higher global scores for tablet-based evaluations over paper-and-pencil-based evaluations is likely due to several factors related to each format. For example, one format may be easier to navigate than the other, facilitating moving back and forth between evaluation items or finding them in real time as behavior occurs. The tablet may have been faster to use than writing by hand, allowing the evaluators to spend more time watching the procedure and increasing the likelihood of observing targeted criteria. Another possibility is that one format may make it easier or harder to use prior items as a point of reference for other items, which would affect consistency within the rating form and potentially affect outcome scores. Although most evaluation processes in research or the real world are likely to use only one method of evaluation, my findings indicate that the selection of method can influence scores unintentionally. I recommend that evaluators give serious thought to what type of evaluation (paper and pencil, electronic, etc.) best suits their purposes and that the same method be used across all trainees once the decision is made.

7.3.2 Suggestions to avoid unintended effects of evaluator cognitive processes on ratings.

Although the circumstances of an evaluation (and thus contextual effects on scores) will vary across tasks and scenarios, influences due to the evaluators themselves will be expected in any situation. Analysis of the demographic data offered several reasons to be encouraged by the performance of the evaluators in the Maryland study. Outcome scores did not vary systematically depending upon the individual person conducting the evaluation (i.e., the identity of the evaluator did not predict scores), the order of the questions on the rating form appeared to have no impact, and I did not observe any sex, domain experience, or other demographic effects that would lead me to suspect systematic bias in the way that evaluators observed the surgeons.

Reliability indices offered a slightly different picture. I used the CWS index to examine the consistency of the small group of evaluators who provided repeated measures data, and found evidence that the evaluators varied in their internal consistency regarding how they applied the scoring criteria, differing both from one another and over time. CWS scores also indicated that individual evaluators were more internally consistent in judging good performers than bad performers. In contrast to the results of Chapter 3, which indicated the reliability of the overall data set, examination of more specific aspects of the data set using Krippendorff's alpha indicated that raters were more likely to disagree at the high and low ends of performance, and that the evaluators demonstrated better interrater agreement as they gained experience.

CWS scores indicate that individual evaluators know good performance when they see it, but are less able to differentiate bad performance (or at least are less able to judge it in a consistent manner). This may be due to the lack of definitions of bad performance – good performance is explicitly defined by the performance measures, but bad performance is mostly defined by the absence of good performance, or the presence of error. Further, there are fewer ways to be "good" (because a high percentage of items must receive high scores) than "bad" (because many types of error may occur or any subsample of items may receive lower scores). Meanwhile, Krippendorff's alphas indicate that criteria for good and bad performance appear to vary across the group of evaluators. Based on these findings, I recommend that any evaluators be trained extensively on the scoring criteria and that they be given a chance to practice using the evaluation form with a variety of trainee skill levels prior to evaluating trainees on their own.

7.3.3 Limitations.

I used similarly-performing surgeons to approximate repeated stimuli when generating CWS scores. While I realize that surgeons who receive similar global scores may differ on specific evaluation items, I reasoned that this decision was justified as more representative of real-world performance evaluation (where evaluators rarely get to observe the same person multiple times in the same circumstances). Our use of aggregated Maryland predictors rather than individual evaluation items should also help to mitigate this issue somewhat.

My calculation of Krippendorff's alpha within performance tiers may have been affected by range restriction. The raters clearly agreed well enough to support aggregating the surgeons into performance bins (i.e., they generally agreed the surgeon did well or did poorly). However, by truncating the range of scores under consideration within a given performance tier, observed differences between scores may have been exaggerated compared to the range of scores in the full data set. Further, my CWS results are based on only three evaluators. A larger sample is clearly needed to draw more firm conclusions about differences between references.

7.3.4 Conclusions.

Although subjective global scores are certainly useful, my analyses indicate that evaluators' judgments are not isolated to the particular case under consideration and that the medium used to conduct evaluations (paper or electronic) can affect global scores. Further, the consistency of evaluations appears to vary both within and between evaluators based on performance level and the experience of the evaluator. Interrater reliability for the data set as a whole does not necessarily indicate reliability within individual parts of the data set.

Though this finding threatens to undermine my other analyses that rested on supposedly reliable measures to identify differences between experts and novices, I do not believe my other findings are in jeopardy. In addition to the limitation of my new Krippendorff's calculations noted above, global assessments were related to the most objective performance metric I could generate with the available data (Chapter 3), indicating a level of predictive validity. The global scores also responded to training interventions designed to improve surgical skill, further supporting the assumption that they capture performance. Finally, the subjective nature of the global score allows an opportunity to capture behaviors that are difficult to articulate in current performance measures, evidenced by the relation of my WSU factors to scores but relatively few Maryland predictors accounting for variance in global scores. This opportunity should not be sacrificed lightly. Global assessments or other subjective ratings still offer valid insight into performance, but acknowledging certain vulnerabilities in scoring can help strengthen performance evaluation metrics to better capture skill.

I note bright spots in the examination of whether factors beyond the surgeons' skill affected performance scores. For instance, I did not find strong evidence that the order of items in the rating form influenced scores, or that systematic demographic biases existed among the evaluators. Nonetheless, I noted several extraneous influences on global performance scores that highlight the potential influence of both the evaluation setting and the evaluators. Traditional notions of rater reliability based on a data set as a whole may miss differences in evaluators over time and across different subsections of the data set. Capturing this nuance may facilitate identification of expert behavior through comparison with novices/journeymen. Though knowledgeable human evaluators remain useful and, indeed, critical to capturing many of the behaviors related to goal establishment and goal enactment, they are not immune from normal cognitive processes. Contextual and cognitive influences on evaluators' judgments must both be accounted for in order to better isolate and measure expert performance.

Chapter 8 – General Discussion

I sought to characterize and expand the range of behaviors used to identify expertise in order to improve measures of performance in the context of a surgical training study. I analyzed archival performance data gathered from resident physicians before and after a training intervention, as well as from experienced attending surgeons. The parent study provided demographic, self-confidence, and performance ratings, to which I added data derived from video and think-aloud protocols using a coding scheme designed to capture various aspects of goal establishment and goal enactment. I utilized the constructs of goal establishment and goal enactment as a guiding framework to inform my sampling of relevant behaviors. In doing so, I identified generalizable constructs not typically included in more task-specific performance appraisals. My analyses examined several aspects of performance and performance measurement, including behaviors that contribute to performance and the role of self-awareness and experience in surgical skill. In addition, I examined the contributions of contextual factors and evaluator cognition to outcome scores in the hopes of generating a holistic view of performance measurement. Due to the descriptive nature of my findings and the subjective nature of my constructs, these findings are best viewed as preliminary results to drive hypothesis generation and future testing. Nonetheless, I believe the results make several contributions to the broader literature.

8.1 Contributions (Theoretical and Methodological)

This study compared residents to attending surgeons. Though residents are not licensed to practice independently, they are not novices in the truest sense of the word (i.e., they have medical training and at least some experience under the guidance of attending surgeons). My results are therefore more accurately described as a journeyman-expert comparison rather than an expert-novice comparison. Nonetheless, my findings speak to several aspects relevant to the conceptualization and measurement of skilled performance among experts.

Chapter 3 identified five factors, motivated by the constructs of goal establishment (identifying problems in the world) and goal enactment (solving problems in context). Goal establishment was represented by the *monitoring* factor, whereas goal enactment was represented by the *instrument change, strategy, deliberate behavior,* and *oddities* factors. Chapter 4 demonstrated that these generalizable factors capture variance in global performance measures, mediated by their relationship to task-specific assessment items. This finding provides encouragement that goal establishment and goal enactment can motivate more generalized domain sampling. Chapter 5 demonstrated that surgeons are aware of their performance, and possibly utilize this awareness to calibrate their level of self-monitoring during a procedure. Chapters 4 and 5 together speak to an interaction between automatic and deliberative processes absent in the expertise literature and extending the cognitive literature into skilled performance. Chapter 6 indicated that raw career experience alone is not the best predictor of surgical skill, challenging a normative definition of expertise in the surgical literature. Finally, Chapter 7 illustrated that factors beyond the skill of the surgeon influence performance scores and must be taken into account when assessing surgical skill, suggesting best practices in performance evaluation that may need to be implemented. Taken together, the findings of the previous chapters indicate that deliberate processes contribute to skilled performance, build a case for new ways of sampling domains, indicate that raw experience is not sufficient for expertise, and indicate that factors beyond the skill of the surgeon contribute to subjective performance measures.

8.1.1 Theoretical contributions.

Psychological theory most often conceptualizes skilled psychomotor performance and expert decision making as relying on unconscious or associative processes, without the need for deliberation or explicit guidance from the conscious mind (e.g., the distinction between declarative and procedural knowledge in cognitive architectures or Klein's (1989) RPD model of expert reasoning). Indeed, I found in Chapter 4 that scores on the *monitoring, strategy,* and *instrument change* factors were negatively related to performance, indicating that the best surgeons completed the task quickly and efficiently, with little need for correction along the way.

At first blush, this finding supports the notion of skilled performance as largely automatic. However, my finding that the *monitoring* factor accounted for variance in performance indicates a cognitively penetrable component to expertise. This construct arose from the verbal protocol data, further reinforcing its conscious accessibility. I found in Chapter 5 that surgeons across levels of skill and experience were aware of their relative skill and were able to adapt their self-perceptions in response to information from the world. I also found that more experienced attending surgeons tended to show more *deliberate behaviors* (specifically naming structures) than the post-ASSET residents.

The expertise literature largely treats automatic and deliberative processing as separate. Both Rasumssen's (1994) SRK framework and Klein's RPD model assume that people only rely on deliberative processes when more automatic processes are inadequate. Interactions between the two are not addressed. Dual-process theories in the cognitive literature have incorporated this interaction, but only in the context of decision making (not skilled behavior). My findings align with the monitoring role of deliberate processes proposed in the dual process literature (Ferreira et al., 2016; Pennycook, Fugelsang, & Koehler, 2015), extending these findings from problem solving into skilled behavior and incorporating the idea of interaction in the expertise literature.

The constructs of goal establishment and goal enactment serve not as newly identified cognitive processes, but as a functionally descriptive framework to help guide domain sampling and ensure that relevant behaviors are captured in measures of performance. Goal establishment and goal enactment contribute to more generalizable theories of expertise to help identify consistency across tasks and domains. Simply because a function (in this case goal establishment and/or goal enactment) results from more basic mechanisms does not relegate it to epiphenomenal status. The function and the mechanism are each crucial to understanding the other (Juvina, 2011). Skilled action does not occur as an isolated task; such behavior occurs within the broader context of a goal-oriented sociocultural system. Just as Clancey argues that systems analysis must account for both the function (the “why”) and action (the “what” or “how) of various components, I argue that both goal establishment (determining the “why” of actions) and goal enactment (the “how”) behaviors are required to successfully perform tasks in the world.

Knowing relevant structures for a given procedure facilitates working efficiently by avoiding false pathways and potentially allows adaptation to unique patient anatomy.

The contribution of the *monitoring* factor to global scores and apparent self-awareness of skill on the part of the surgeons (goal establishment), along with experience differences in components of the *deliberate behavior* factor (goal enactment) are all consistent with the view that skilled behavior has a cognitively penetrable component, allowing experts to perform their tasks within their work domain.

8.1.2 Methodological contributions.

Measures of skilled performance suffer from difficulty in identifying relevant domain behaviors that generalize beyond the task at hand. As a result, current surgical performance measures are highly procedure-specific and labor-intensive to create. Further, surgical research often pre-identifies experts based partly on experience rather than demonstrably superior outcomes. The findings of my study offer an additional perspective to incorporate into existing medical views of expert performance, provide a guide for how to sample a domain to improve generalizability and resolution, and also offer possible insight into weaknesses of current measures of interrater reliability. Finally, I offer a word of caution to those in the medical community who weight experience heavily in operationalizing expertise.

8.1.2.1 Goal establishment and goal enactment guide sampling of domain behaviors.

I found in Chapter 4 that components of goal establishment (the *monitoring* factor) and goal enactment (the *instrument change* and *strategy* factors) predicted both Maryland global scores and the more generalized Maryland adjusted TRI and fasciotomy IPS measures. Although the Maryland predictors (particularly those related to technique) fared better in capturing variance in these scores, I believe this is due to the fact that these items are specifically shared across procedures. On the other hand, my WSU factors representing goal establishment and goal enactment captured variance in outcome scores across procedures despite not being tied directly to the items under consideration.

Goal establishment and goal enactment guide domain sampling to allow us to strike a balance between overly specific content analyses and overly general cognitive analyses by focusing on how people identify and solve problems in the work environment. Goal establishment is particularly important to capture, as laboratory-based studies of expertise tend to omit how people parse the world in ill-structured domains (e.g., Patel). Current task-specific evaluation methods place too much emphasis on raw technique. Proper technique is certainly important to skilled performance, but such a focus omits key features of expertise in the broader work domain. The Maryland variables capture the "what" of behavior in the form of technique; my WSU factors capture the "why" as well (particularly the monitoring factor). Together these measures point to basic competencies applicable across situations, but not captured by current surgical metrics. The ability to parse the world and monitor progress are key components of expert behavior that to date have not been included in surgical performance measures.

8.1.2.2 Current measures of interrater reliability do not account for nuance in the data set.

I found that evaluator consistency and reliability (as measured by CWS scores and Krippendorff's alpha) varied across evaluators as well as across levels of surgical skill and evaluators' experience judging performance. Although the evaluations as a whole were reliable, and global performance scores appear to be a valid measure of performance, these findings indicate that evaluators do not judge all observations equally. I found evidence that evaluators do not treat observations independently, and that factors beyond the skill of the surgeon can impact performance ratings. These external influences must be accounted for in order to better isolate (and thus measure and understand) skilled performance.

8.1.2.3 Process is more useful than experience in identifying experts.

Expertise theory and studies that seek to operationalize expertise often assume that developing expertise requires time to amass enough deliberate practice (Ericsson, Krampe, & Tesch-Romer, 1993) or develop a sufficient repertoire of patterns (Klein, 1989; Simon & Gilmartin, 1973). In opposition to the common view that expertise requires many years to develop, (and consistent with current medical performance measures' difficulty distinguishing among high performers) I found evidence that some residents were capable of performing at a level comparable to the attending surgeons (and that some attending surgeons performed similarly to residents). I also found that post-ASSET residents and attending surgeons received statistically similar Maryland global performance scores. These findings indicate that experience is not the best way to identify experts. Process-based measures, however, provide useful insight into skill. I found that the best-performing surgeons (resident or attending) displayed nearly identical scores on all predictors of performance, with the exception that the attending surgeons were faster. When compared to the post-ASSET residents as a group, however, I found that the attending surgeons had better scores on measures of technique and demonstrated lower scores on the *instrument change* and *strategy* factors. These results suggest that process-based measures such as those incorporating goal establishment and goal enactment are a far more effective way to identify expertise than experience or even outcome-based measures.

8.1.2.4 Think-aloud protocols tend to capture negative behaviors.

One of the issues of measurement highlighted by negative relationships between WSU predictors and performance is that think-aloud protocols tend to capture negative behaviors. Behaviors such as “expressing doubt or uncertainty” are easier for a layperson to identify and negative thoughts are generally more likely to be articulated in a think-aloud protocol (e.g., a surgeon may articulate a plan, but has no inherent reason to say anything further when the plan unfolds correctly. The surgeon can implicitly acknowledge success simply by continuing on when plan execution has been completed without further modification. If something goes wrong, on the other hand, the surgeon must acknowledge it and generate a new course of action.). However, by capturing largely negatively-valenced behaviors, the relationship between behavior and performance in think-aloud data may be altered. This issue is discussed further in both the limitations and future research sections.

8.2 Limitations

My analyses were based on archival data derived from a surgical training study conducted at the University of Maryland to improve trauma surgeons’ readiness to perform rare procedures. With any archival study, there are certain disadvantages that come with using existing data for a new purpose. Many of the limitations of the present work are a byproduct of my reliance on archival data and quirks in the design of the Maryland study. These limitations include the nature of my think-aloud data, aspects of the comparison between residents and attending surgeons, the rarity of the procedure selected for examination, and the aforementioned sampling issues. My own analysis decisions may also have affected results. I start with an overview of the Maryland study design to help frame these issues, followed by discussions of each of the identified limitations.

8.2.1 Design overview.

Archival Maryland data consisted of demographic data, self-reported confidence, and performance evaluations based on multiple evaluators' judgments of cadaver- and model-based procedures before and after training. I derived new variables based on audio transcripts and video of the procedures in order to capture cognitive aspects of performance.

Attending surgeons participated along with resident surgeons, leading to a somewhat unique design whereby the residents received a within-subjects manipulation to be compared to a separate group of surgeons (the attending surgeons). Time and participant recruiting constraints in the parent Maryland study necessitated that attending surgeons, pre-training residents, and post-training residents were evaluated as groups, with the attending surgeons first, followed by the pre- and post-training residents, respectively. As a result, the experience of the evaluators was confounded with the group under evaluation.

8.2.2 Think-aloud data may have been incomplete.

Among the most prominent limitations of my study is the use of think-aloud data. As with any study involving think-aloud data, I cannot be certain that the surgeons articulated everything that they were thinking - even under the best of circumstances I would be unable to guarantee that a surgeon did not consider something simply because it was not articulated aloud. This issue is particularly relevant for the present study. The original study from which my data came was designed to assess training effectiveness rather than expert-novice differences in reasoning. The instructions and prompts provided to participants were oriented towards identifying knowledge of anatomy and procedural steps rather than cognitive strategies per se. Different instructions or prompts may have yielded different content (such as more positively-valenced behaviors) and a different pattern of results.

8.2.3 Comparing residents to attending surgeons introduces confounds.

One possible confound in this study (largely unavoidable due to the nature of surgical training) is the comparison of residents to attending surgeons. The resident surgeons were not yet specialized, while the attending physicians were all specialized trauma or thoracic surgeons. Time on task and the nature of practice are therefore confounded. Some of the observed experience-related performance differences could have been due to the fact that the residents did not have the same focused practice as the attending physicians rather than due to broader differences in knowledge or technical ability. Given that some residents performed comparably to the best attending surgeons and that some attending surgeons performed poorly, I do not believe experience alone to be the best predictor of performance.

In a more general sense, including journeymen rather than true novices in the sample may have affected my findings by altering the comparisons I was able to make. For instance, it is possible that had I used true novices (such as a medical student who has completed a surgical rotation but lack extensive experience), the *monitoring* factor would have shown an increase between this novice group and residents. However, in a domain such as medicine a true novice with no training or experience does not always provide useful insight (i.e., a layman lacks any structure or skill – guessing behavior does not provide an informative contrast). Residents (journeymen) display enough structured behavior to identify features characteristic of expertise while still offering a reasonable skill contrast to the attending surgeons (experts).

8.2.4 The rarity of the studied procedure may have influenced surgeon behavior.

Another possible issue is that this study used procedures that are rare in daily practice, meaning that I studied nonroutine expertise likely involving more thought or problem solving than required for more common procedures. This may have led to different findings than had I examined more frequently performed procedures. For example, many residents stated that they had never completed an axillary artery exposure before. These residents likely had to problem solve and use different thought processes than the surgeons with better anatomical knowledge or who had performed the procedure before. The gap in performance between experts and novices or the nature of performance differences likely would have been different had I examined more familiar procedures.

8.2.5 Small samples and family error may have affected my findings.

As described in Chapters 5, 6, and 7, sampling issues due to missing data or due to analysis of subsamples of the data set could have led to spurious findings and/or reduced power. Sample sizes for many analyses were small, precluding strong conclusions. Further, the sheer number of analyses coupled with those same sampling issues demand caution in interpreting any single finding. Any positive result could be due to sampling error, or a negative result could be due to lack of power. My results in this study are therefore best considered preliminary or suggestive. Nevertheless, I think my findings taken as a whole form a coherent narrative with potentially important implications for performance measurement and should be explored further. We need a starting point to begin to frame our explorations of expert behavior. The parent Maryland study utilized one of the largest samples of representative surgical behavior available. Other studies have used part-task or laparoscopic stimuli. This study is a step towards a generalizable theory of expertise to determine whether the findings from those studies generalize to other tasks (e.g., from laparoscopic to open surgery).

8.2.6 My decisions during the analysis may have affected the trends I was able to identify.

In addition to the limitations described in previous chapters (e.g., only analyzing a small subsample of the data set), other analysis decisions may have affected my results. The necessity of investigating new questions with a data set designed for a particular purpose required me to make certain decisions related to the analysis that, although justifiable, were not necessarily the only possible option. For instance, the original University of Maryland study utilized both surgical models and cadavers in the post-ASSET evaluations, causing some of the cadaver-based assessments to be abridged by omitting the knowledge-based evaluation items. In my analyses, I combined scores on the knowledge-based items from the evaluation using the surgical model with the procedural items from the evaluation using the cadaver. This decision increased my available sample size, but somewhat reduced my ability to make apples-to-apples comparisons between pre- and post-ASSET procedures. As another example, multiple evaluators provided quantitative performance ratings in the original University of Maryland study, but only a single rater provided evaluations of the video and think-aloud data for the present study. This necessitated paring down the quantitative data in order to provide a more one-to-one match between predictors from the original study and the new predictors added in the current work. Combining Maryland raters came with certain tradeoffs (such as not being able to utilize multilevel modeling in most analyses), but I felt that combining the quantitative ratings from each individual Maryland rater was the best approach to this problem for my purposes. Other approaches could certainly have been justified as well and may have provided different insights into the data set.

Further, I deliberately overlooked the repeated nature of the pre- and post-ASSET evaluations for certain analyses. While this decision increased the available range of experience and performance in the sample, it also limited power and may have introduced spurious relationships in the data. By ignoring these correlations between some procedures I potentially altered the observed relationship between predictors and outcome, as well as the observed relationships among the predictors themselves. I computed an ICC for procedures nested within the resident surgeons as a means to judge how much variance in global scores was attributable to the individual surgeons (and hence how likely these correlations would be to affect my findings). The ICC was zero, indicating that none of the variance in global scores was attributable to the individual surgeon. I am skeptical of this result due to the small number of procedures within each resident (two) and the presence of a warning in the output that I was unable to resolve. However, given the impact of the ASSET course on global scores and the statistical similarity between post-ASSET residents and attending physicians on global scores, it seems plausible to believe that the training intervention at the individual procedure level within each surgeon accounted for a much greater proportion of the variance in outcome scores than individual differences between the surgeons. I therefore feel reasonably confident that the repeated measures aspect of the study had little, if any, effect on the results.

8.3 Future Research

Though this study has added to the expertise literature (by incorporating interactions between automatic and deliberate processes) and the cognitive literature (by extending these interactions into skilled behavior), many more questions remain unanswered. Here I address remaining theoretical questions in the areas of replicating my findings**,** exploring interactions among automatic and deliberate processes, and skill development, as well as practical questions of how best to capture expertise.

8.3.1 Replicating my findings in other domains will increase our understanding of shared requirements across tasks and domains.

Future work should apply this approach to the three remaining Maryland procedures in order to explore whether the same types of constructs apply to other procedures within the surgical domain. Finding similar constructs would strengthen my argument that the *monitoring, instrument change, strategy, deliberate behavior,* and *oddities* factors are generalizable and useful for evaluations within a domain rather than on a per-task basis. I would also like to replicate this method in other domains. Successful identification of higher-level constructs beyond task-specific actions in other domains would serve as a useful replication of my findings. The results of such an analysis in other domains would also be informative to theories of expertise and help us to improve our understanding of the necessary skills in various types of tasks. To the extent that the same types of processes occur across tasks and domains we can make stronger assertions about the broader role (or lack thereof) of cognition and deliberate processes in expert behavior as well as begin to identify “types” of domains that may share common skills and cognitive requirements.

8.3.2 The nature of the interaction between automatic and deliberate processes needs to be clarified.

The cognitive dual process literature has framed the interaction between automatic and cognitive processes as linear (heuristic decision processes are activated by the environment, which are then double checked and overridden if needed by supposedly superior deliberate thought; Ferreira et al 2016). This line of thinking assumes that heuristic-type processing is flawed and analytical reasoning is always correct, but that is not necessarily the case (Gigerenzer, 2008). Future research should identify how the decision making process handles instances in which conflict arises due to flaws in analytical rather than heuristic thought.

Further, the influence of conscious processes on automatic processes has only been discussed on the back end of problem solving (inhibiting heuristic output, for instance). For example, Ferreira et al. (2016) argue that heuristic problem solving is triggered by the features of the problem and that goals or intentions should not have an impact. However, experts see different features of problems than novices, and filter information differently (Chi, Feltovich, & Glaser, 1981). Depending on the goals of the expert, or on how conscious processes influence the parsing of the world (via knowledge or attention), high level thought may still affect heuristic decision making by altering which features of the problem become salient. Influences of automatic and deliberate processes on one another should be considered throughout the problem solving process.

8.3.3 We must clarify the role of experience and training in skill acquisition.

In addition, we should seek to identify the factors that facilitated expert-level performance in some of the less-experienced resident surgeons. These surgeons excelled at the procedure after relatively brief training and in the absence of extensive general surgical experience (and even less, if any, experience with the particular procedure in question). We should work to identify what experiences, traits, and cognitive processes facilitated such high performance in the absence of extended practice, or what skills carried over from more practiced procedures. Expertise and training theory would benefit from a more precise understanding of which skills come with experience vs. training, and any role of certain selectable traits in such skill acquisition. Perhaps we can better design training courses to allow surgeons to rapidly reach expert levels of performance and adapt to novel procedures more effectively. Alternatively, we can work to identify surgeons who are likely to learn quickly or may need additional training to reach asymptote without years of experience.

8.3.4 We must work to identify affirmative indicators of expertise.

We should make an effort to identify positive behaviors as indicators of cognition as well as negative behaviors. For example, the nature of the behaviors within the *Monitoring* factor may have affected the relationship between this factor and performance. The *monitoring* factor included the negatively-valenced behaviors of expressing doubt and recognizing a mistake, but not positively-valenced behaviors. By making more of an effort to identify affirmative behaviors associated with expert performance, we can start to use this technique to study experts based on what they do rather than what they avoid doing. We can also gain a better sense of the true contribution of metacognitive or deliberate processes to expertise.

8.4 Conclusions

This study utilized archival data from a trauma surgical training study to examine the cognitive aspects of expert performance and work towards more generalized performance measures within a domain. I identified five higher-level factors related to goal establishment (the *monitoring* factor) and goal enactment (the i*nstrument changes,* *strategy, deliberate behavior,* and *oddities*) related to surgical performance. Goal establishment and goal enactment serve as a useful framework to guide domain sampling and expand the identified range of behaviors that should be captured by performance measures in the process. I have made a case that task-level performance can be captured with broader domain-relevant constructs, and that the traditional technically-focused view of performance assessment should be expanded to include more cognitively oriented constructs. I have also called into question the experience-dominated view of expertise within the medical community, and highlighted the need to consider the influence of environmental and contextual features on outcome measures. These findings can be used to generate hypotheses to begin work towards a more complete (yet generalizable) view of expert level performance.

APPENDIX A

IN-PERSON EVALUATION SHEET/SCRIPT

Name of Evaluator: **Date:**

Name of Candidate: **(Circle timing): Pre Post**

1st Trial

Circle type of trial: Cadaver / Model

Case One: Axillary Artery

Case Presentation:

- **You are called to the Emergency Department to see a 24 y/o male who was shot during an attempted robbery sustaining a single gunshot wound to the upper anterior lateral Right/Left Chest.**
- **He was reported to have a large amount of bright red blood at the scene, but is currently not bleeding.**
- **He is complaining of pain at the site of the wound and inability to move his arm.**

[Advance slide to show image of wound]
[Advance slide to continue narrative]

- **He is awake and talking with bilateral and equal breath sounds and a BP of 80/60 and a heart rate of 130 after 2 liters of lactated ringers**
- **There is a single wound as seen with no other obvious trauma and no "exit wound". His hand is cool and pale.**

Q1: Question #1. What are the structures you suspect could be injured along the path of the bullet?

Expected Answers checklist:

S1: The participant described each of the following as potentially injured:		
	Yes	**No**
A1: Axillary Artery		
Axillary Vein		
Brachial Plexus		
Lung		
Subclavian Artery		
Subclavian Vein		
Mediastinal structures		
A8: Bones		

Q2: Question #2. What physical findings will you look for to help you decide which structures are injured? Include signs of vascular, thoracic, nerve, and bone injury.

Expected Answers checklist:

S1: The participant describes each of the following physical findings and tests:		
	Yes	**No**
A1: Decreased breath sounds		
Active arterial bleeding		
Enlarging or expanding Hematoma		
Absent distal pulses		
Distal Ischemia		
Bruit or palpable thrill		
- Indicates that any or all of above are "hard signs" of vascular injury		
Active venous bleeding		
Unequal blood pressure, decreased Brachial-Brachial Index		
Doppler pulses—diminished flow		
Sensory loss		
Loss of motor function – weakness, inability to move arm		
Bony instability, deformation, crepitus		
Sub-cutaneous air		
A15: Tracheal deviation		

The patient's blood pressure is 85/65 and HR 110 and is unable to move his arm, has decreased sensation and absent brachial, radial, and ulnar pulses.

**Q3: Question #3:
What additional studies would you perform to help you identify or rule out specific injuries in this patient?**

Expected Answers checklist:

S1: The participant described each of the following as additional studies		
	Yes	**No**
A1: FAST exam to look for pericardial tamponade, hemothorax, pneumothorax		
Chest X-ray		
A3: A marker (eg paperclip) is placed to mark wound prior to x-ray		
E1: Error: Fails to obtain CXR		
A4: CT of Chest (zero points)*		
CT Angiogram (zero pts)*		
A6: Angiogram (zero points)*		
E2**:** Error: Inappropriate use of CT or Angio*		

****All of the above tests are acceptable possible studies but the participant should clearly indicate these tests <u>should only be done in a hemodynamically stable patient</u>. Without this qualifier, performing any of these tests prior to taking this patient to the OR has potential for negative outcome & should result in negative value scoring.***

***Scoring Note: no additional points are added for additional studies**

[Advance slide to show Chest x-ray]

A chest x-ray has been obtained and shows no evidence of hemo or pneumothorax. There is a bullet fragment adjacent to the mid-portion of the ipsilateral scapula just superficial to the skin of the back – In other words a bullet trajectory from front to back on the same side, which does NOT involve the thoracic cavity.

Now the BP is 89/69 HR is 110. There is no other obvious trauma and his hand is cool and pale.

Q4: Question #4:
Now that you have seen the wound, physical findings, and chest x-ray, what is your plan for this patient?

If the participant suggests a non-operative course – they should be informed that: **the patient is now in the operating room and needs exposure and control of the axillary artery.**

Expected Answers checklist:

S1: The participant states the following plan		
	Yes	No
A1: Patient should be taken urgently to the Operating room		
E1: Error: Delay in going to the operating room		

Q5: Question #5: What is your plan to resuscitate this patient? Include fluids or medications you would use during the initial resuscitation.

Expected Answers checklist:

S1: The participant describes each of the following additional items the patient might receive:		
	Yes	No
A1: Resuscitate with blood products		
Transfuse with high ratio of blood:FFP:platelets/ Massive transfusion protocol		
Minimize crystalloid infusion		
Limit volume resuscitation until bleeding controlled		
Do not delay surgery for resuscitation, resuscitate in OR		
Give TXA		
A7: Large bore IV access		
The patient has now been transported to the Operating Room and is on the OR table in front of you.		

Question OR # 1: (Q6) How would you position and prep this patient in order to repair this injury and explain why you chose to prep as you did?

Expected Answers checklist:

S1: The participant Indicates the following in response:		
	Yes	No
A1: The patient should be supine		
A2: The arm extended on an arm board		

S2: The prep should include:		
A1: The Entire Chest		
States possible need for sternotomy for proximal control		
The Entire arm and hand on the affected side		
States need to evaluate perfusion to the hand		
The thigh/groin for possible vein harvest		
The neck		
States possible need to expose subclavian artery for proximal control		
S2E1: Error: Fails to prep entire chest		
S2E2: Error: Fails to prep entire arm and hand.		
S2E3: Error: Fails to prep the thigh for vein harvest		

Question OR # 2: Q7
At this time, please describe and then mark on the skin the landmarks and the incision that you plan to use.
Expected Answers checklist:

S1The participant Indicates the following in response:		
	Yes	**No**
S1A1The sternal notch		
The clavicle		
The deltopectoral groove		
S1A4:Incision runs from mid-clavicle laterally in deltopectoral groove.		

EXPOSURE OF AXILLARY ARTERY

"Now I would like you to get control of the Axillary Artery proximal to the wound by dissecting and placing a vessel loop around the artery. As you operate, speak out loud and identify each step of the procedure. It is not necessary to rush through the procedure—you should operate at a comfortable pace. The procedure will be deemed complete once you have placed a vessel loop around the axillary artery to obtain proximal control. Do you have any questions? If not please proceed."

Q8: Expected operative dissection performance checklist:

The participant describes and performs each of the following steps:			
	Yes	**No**	**Time**
S1A1: Initial skin incision is adequate to perform exposure			**Start Incision** **Blank**
Splitting or dividing Pectoralis Major			**Start Dissection** **Blank**
Divides Pectoralis Minor			
Correctly identifies Axillary Artery			
Correctly identifies Axillary Vein			
Correctly identifies brachial plexus			
S1A7: Controls the Axillary Artery Proximal to injury			**Finish** **Blank**
S1E1: Error: Incorrectly identifies the Axillary artery and does not recognize or correct error		**Q8S1A7_time:** *(indicates duration of procedure)*	
S1E2: Error: Incorrectly identifies the Axillary Artery but is able to recognize and correct			

***Q8S2:* Technique points**

	Score 1-5
Q8S2A1: Exposes arteries by dissecting directly on anterior surface†*	
Manipulates artery by grasping adventitia†*	
Uses instruments properly	
Positions body to use instruments to best advantage	
Proceeds at appropriate pace with economy of movement	
Handles tissue well with minimal damage	
Creates an adequate visual field using retractors for procedure	
Communicates clearly and consistently	
Performs procedure without unnecessary dissection	
Q8S2A10: Continually progresses towards the end goal	

(5) Every time/Excellent; (4) Almost every time/Very good; (3) Sometimes/Good; (2) Rarely/Fair; (1) Never/Poor

*N/A for model, ***†Score (1) if participant never finds an artery***

Q9S1: Expert Discriminator Operative Field Maneuvers for Axillary Artery Exposure

	Yes	**No**
Q9S1A1: Operates through 'key-hole' or too small a skin incision	0	1
Operates using full incision	1	0
Excessive dissection	0	1
Pointless digging and shifting around in surgical field	1	0
Has a logical operating sequence	1	0
Q9S1A6: Lacks anatomical knowledge	0	1

Q9S2 : Expert Discriminatory Instrument Use for Axillary Artery Exposure

	Yes	**No**
Q10S1A1: Improper instrument use (e.g. back-handed use)	0	1
Incorrect instrument holding (e.g. forceps too near tips, thumb through scissors handle)	0	1
Scalpel use: multiple tentative cuts or cuts tangentially	0	1
Switches instruments excessively	0	1
Effective use of blunt dissection	1	0
Dedicated use of a single instrument.	0	1
Q10S1A7: Uses sharp dissection (knife or scissors) confidently	1	0

Questions in OR, after dissection:

Q11S1: What are the consequences of ligating the axillary artery?

The participant answered the questions correctly:		
	Yes	**No**
A1: Ligation of the axillary generally does not cause ischemia due to extensive collaterals around the shoulder.		

Q12S1: What are the pitfalls or common errors that one might expect with this procedure?

Possible Answers		
	Yes	**No**
A1: Incision – too high, too low, wrong location		
Iatrogenic injury to nerve, artery, vein		
Inability to get proximal control – needing to go above clavicle or into chest		
Diving into clot or hematoma without adequate control		
A5: Mistaking nerve for artery		

AXILLARY ARTERY EXPOSURE GLOBAL RATING (circle one)**:**

G1: Overall Understanding of the Evaluation and Treatment of a Patient with a Suspected Axillary Artery Injury:

1	**2**	**3**	**4**	**5**
Core knowledge is poor and there is no evidence of understanding the nuances of evaluation and diagnosis.	Core knowledge is fair with some understanding of the nuances of evaluation and diagnosis.	Core knowledge is good with moderate understanding of the nuances of evaluation and diagnosis.	Core knowledge is very good with thorough understanding of the nuances of evaluation and diagnosis.	Core knowledge is excellent with a superior understanding of the nuances of evaluation and diagnosis.

G2: Overall Understanding of the Surgical Anatomy of the Axillary Region:

1	**2**	**3**	**4**	**5**
Poor knowledge of the regional anatomy. Unable to identify major structures or their relationships.	Fair knowledge of regional anatomy. Can name some of the major structures and their relationships	Good understanding of the anatomy. Can name most of the major structures and their relationships.	Very good understanding of anatomy. Able to point out all of the major structures and their relationships.	Excellent understanding of the anatomy, including variants. Knows the minutia, Should be teaching anatomy class.

G3: Technical Skills for Exposing the Axillary Artery:

1	**2**	**3**	**4**	**5**
The participant's technical skills were poor with much wasted moves and very poor tissue handling.	The participant demonstrated fair technical skills with some wasted movements and errors in tissue handling	The participant demonstrated good technical skills with occasional wasted movements and errors in tissue handling.	The participant demonstrated very good technical skills with minimal wasted movements and errors in tissue handling.	The participant demonstrated excellent technical skills with no wasted movements and proper respect for tissues.

G4: This participant is ready to perform exposure and control the Axillary Artery:

1	**2**	**3**	**4**	**5**
The patient has exsanguinated. Participant is not ready to perform the exposure.	This participant could do the exposure fine with experienced help, but will struggle if left alone.	The participant might need to look at a text to refresh their memory but will be able to perform the exposure.	This individual will be able to perform the exposure with minimal difficulty in an expeditious fashion.	Absolutely, I hope that this individual is on call if I am injured.

ER: Evaluator's overall rating (1-100) ______________________________

≥ 90 Excellent I hope that this individual is on call if I am injured

80-89 This individual will be able to perform the exposure with minimal difficulty in an expeditious fashion.
70-79 The participant might need to look at a text to refresh their memory but will be able to perform the exposure
60-69 This participant could do the exposure with experienced help, but will struggle if left alone
<60 The patient has exsanguinated. Participant is not ready to perform the exposure.

The overall score should be the instructor's subjective rating of how well the surgeon performed. This will be compared to the objective score for the purpose of validating the scoring method.

BH: Body Habitus of cadaver (Circle):

Obese Average Thin

CA: Cadaver Anatomy (Circe):

Normal Variant

APPENDIX B

KRIPPENDORFF'S ALPHA VALUES FOR ALL WSU VARIABLES

Evaluation item	*Alpha*	*95% CI*
Realzes mistake	0.00	-1.00 - 0.00
Realizes lost	0.62	-0.13 - 1.00
Anticipating issues	0.54	-0.19 - 1.00
Forms a plan	0.87	0.64 - 0.99
Weighing options	0.92	0.00 - 1.00
Expresses uncertainty	0.97	0.93 - 1.00
Mentioning things they expect to happen	0.80	0.54 - 1.00
Evaluating progress	0.73	0.47 - 0.91
Risk mitigation	0.62	-0.13 - 1.00
Accounting for individual anatomy	0.94	0.00 - 1.00
Workarounds	0.74	0.00 - 1.00
Facilitation	0.00	-1.00 - 0.00
Naming structures	0.98	0.94 - 1.00
Evaluating structures (appearance)	0.46	-0.27 - 0.91
Evaluating structures (behavior)	0.79	0.00 - 0.79
Evaluating structures (process of elimination)	0.00	-1.00 - 0.00
Evaluating structures (other)	0.00	-1.00 - 0.00
Knowledge	0.62	-0.13 - 1.00
Balancing constraints	0.85	0.00 - 1.00
Evaluator prompting	0.81	0.49 - 1.00
Misc. oddities	0.71	0.27 - 1.00
Environment adjustment (patient)	0.71	0.27 - 1.00
Extending incision	0.88	0.00 - 1.00
Proportion of the time using instruments in two hands (incision)	0.79	0.37 - 1.00
Proportion of the time using instruments in two hands (muscle)	0.73	0.00 - 1.00
Proportion of the time using instruments in two hands (identification)	0.98	0.96 - 1.00
Proportion of the time using instruments in two hands (control)	0.98	0.96 - 0.99
Total instrument changes	0.76	0.39 - 0.98

Instrument changes (incision)	0.65	0.10 - 1.00
Instrument changes (muscle)	0.92	0.79 - 1.00
Instrument changes (identification)	0.90	0.69 - 1.00
Instrument changes (control)	0.91	0.76 - 1.00
Cuts per second for incision	0.54	0.14 - 0.88
Seconds per cut for incision	0.54	0.15 - 0.87
Number of cuts to make incision	0.53	-0.24 - 0.99
Shifts between blunt and sharp dissection	0.98	0.94 - 1.00
Proportion of the time using sharp dissection (incision)	0.83	0.64 - 0.99
Proportion of the time using sharp dissection (muscle)	0.90	0.83 - 0.97
Proportion of the time using sharp dissection (identification)	0.95	0.91 - 0.98
Proportion of the time using sharp dissection (control)	0.80	0.40 - 1.00
Proportion of the time using blunt dissection (incision)	0.33	-0.30 - 0.92
Proportion of the time using blunt dissection (muscle)	0.91	0.76 - 0.99
Proportion of the time using blunt dissection (identification)	0.82	0.57 - 0.99
Proportion of the time using blunt dissection (control)	0.57	-0.27 - 1.00
Exploration	0.76	0.41 - 0.96
Evaluating structures by feel	0.92	0.00 - 1.00
Checking by feel	0.61	0.33 - 0.80
Backtracking	0.78	0.44 - 1.00
Time between identifying the artery and placing the loop	0.99	0.99 - 1.00
Total time (incision)	0.40	-0.76 - 1.00
Total time (muscle)	0.99	0.97 - 1.00
Total time (identification	0.85	0.55 - 1.00
Total time (control)	0.99	0.98 - 1.00
Total completion time	1.00	1.00 - 1.00
Time searching for instruments	0.96	0.93 - 0.99
Idle time	0.68	0.27 - 0.97
Double checking*	1.00	
Evaluating structures (location)*	1.00	
Heuristics*	1.00	
Evaluator hint*	1.00	
Altering a vessel loop*	1.00	
Environment adjustment (workspace)*	1.00	
Environment adjustment (self)*	1.00	

Repositioning retractors*	1.00
Placing retractors*	1.00
Laying out instruments ahead of time*	1.00
Evaluator assistance*	1.00
Expresses confidence*	1.00
Dominant hand*	1.00
Side of the cadaver operated on*	1.00

*Variables demonstrating perfect agreement between the samples do not have a confidence interval.

APPENDIX C

KRIPPENDORFF’S ALPHA VALUES FOR ALL MARYLAND VARIABLES

Evaluation item	*Alpha*	*95% CI*
Cadaver habitus	0.61	0.43 - 0.78
Cadaver anatomy	0.00	-1.00 - 0.48
Q1S1A1: Suspects Axillary Artery injury	0.62	0.35 - 0.89
Q1S1A2: Suspects Axillary Vein injury	0.70	0.53 - 0.85
Q1S1A3: Suspects Brachial Plexus injury	0.68	0.48 - 0.84
Q1S1A4: Suspects lung injury	0.61	0.40 - 0.79
Q1S1A5: Suspects Subclavian Artery injury	0.83	0.69 - 0.94
Q1S1A6: Suspects Subclavian Vein injury	0.70	0.54 - 0.84
Q1S1A7: Suspects mediastinal structure injury	0.60	0.41 - 0.76
Q1S1A8: Suspects injury to bones	0.76	0.62 - 0.88
Q2S1A1: Looks for decreased breath sounds	0.71	0.55 - 0.87
Q2S1A2: Looks for active arterial bleeding	0.68	0.49 - 0.84
Q2S1A3: Looks for enlarging or expanding hematoma	0.71	0.55 - 0.87
Q2S1A4: Looks for absent distal pulses	0.29	-0.10 - 0.68
Q2S1A5: Looks for distal ischemia	0.02	-0.27 - 0.30
Q2S1A6: Looks for bruit or palpable thrill	0.59	0.25 - 0.86
Q2S1A7: Indicates that Q2S1A1-A6 are "hard signs"	0.51	0.14 - 0.82
Q2S1A8: Looks for active venous bleeding	0.32	-0.16 - 0.71
Q2S1A9: Looks for unequal blood pressure	0.25	0.00 - 0.50
Q2S1A10: Looks for Doppler pulses - diminished flow	0.42	0.09 - 0.71
Q2S1A11: Looks for sensory loss	0.38	0.18 - 0.58
Q2S1A12: Looks for loss of motor function	0.58	0.30 - 0.81
Q2S1A13: Looks for bony instability, deformation, or crepitus	0.81	0.67 - 0.93
Q2S1A14: Looks for sub-cutaneous air	0.42	0.09 - 0.71
Q2S1A15: Looks for tracheal deviation	0.34	-0.09 - 0.71
Q3S1A1: Orders a FAST exam	0.80	0.66 - 0.92
Q3S1A2: Orders a chest X-ray	0.84	0.62 - 1.00
Q3S1A3: Marks wound prior to X-ray	0.66	-0.01 - 1.00
Q3S1A4: Orders chest CT	0.66	0.00 - 1.00

Q3S1A5: Orders CT angiogram	0.67	0.44 - 0.91
Q3S1A6: Orders angiogram	0.22	-0.43 - 0.74
Q3S1E1: Error - does not order X-ray	0.78	0.55 - 0.94
Q3S1E2: Error -inappropriate use of CT or angiogram	0.42	0.00 - 0.83
Q4S1A1: Sends patient to operating room	0.39	-0.43 - 1.00
Q4S1E1: Error - delay in going to operating room	0.85	0.41 - 1.00
Q5S1A1: Resuscitates with blood products	0.42	0.03 - 0.74
Q5S1A2: Massive transfusion protocol	0.37	0.16 - 0.56
Q5S1A3: Minimize crystalloid infusion	0.05	-0.28 - 0.38
Q5S1A4: Limit volume resuscitation until bleeding controlled	0.47	0.02 - 0.82
Q5S1A5: No delay - resuscitate in OR	0.12	-0.18 - 0.39
Q5S1A6: Give TXA	0.79	0.47 - 1.00
Q5S1A7: Large bore IV access	0.72	0.56 - 0.87
Q6S1A1: Patient is supine	0.16	-0.48 - 0.68
Q6S1A2: Patient has arm extended on arm board	0.26	-0.17 - 0.63
Q6S2A1: Preps entire chest	0.32	0.02 - 0.62
Q6S2A2: States possible need for sternotomy	0.57	0.37 - 0.77
Q6S2A3: Preps entire arm and hand on affected side	0.23	-0.01 - 0.45
Q6S2A4: States need to evaluate perfusion to the hand	0.42	0.03 - 0.74
Q6S2A5: Preps thigh/groin for possible vein harvest	0.79	0.62 - 0.92
Q6S2A6: Preps the neck	0.39	0.19 - 0.59
Q6S2A7: States possible need to expose subclavian artery	0.01	-0.26 - 0.28
Q6S2E1: Error - fails to prep the entire chest	0.25	-0.10 - 0.60
Q6S2E2: Error - fails to prep the entire arm and hand	0.27	0.02 - 0.52
Q6S2E3: Error - fails to prep the thigh for vein harvest	0.46	0.24 - 0.68
Q7S1A1: Marks the sternal notch as a landmark	0.91	0.81 - 0.98
Q7S1A2: Marks the clavicle as a landmark	0.61	0.40 - 0.82
Q7S1A3: Marks the deltopectoral groove as a landmark	0.70	0.54 - 0.84
Q7S1A4: Marks the incision from the mid-clavicle laterally in the deltopectoral groove	0.54	0.36 - 0.70
Q8S1A1: Initial skin incision is adequate	0.74	0.59 - 0.87
Q8S1A2: Splits or divides Pectoralis Major	0.93	0.83 - 1.00
Q8S1A3: Divides Pectoralis Minor	0.78	0.64 - 0.90
Q8S1A4: Correctly identifies Axillary Artery	0.83	0.68 - 0.94
Q8S1A5: Correctly identifies Axillary Vein	0.65	0.45 - 0.83
Q8S1A6: Correctly identifies Brachial Plexus	0.74	0.59 - 0.87
Q8S1A7: Controls the Axillary Artery proximal to	0.71	0.56 - 0.87

the injury		
Q8S1E1: Error - incorrectly identifies the Axillary Artery and does not recognize or correct it	0.66	0.48 - 0.85
Q8S1E2: Error - incorrectly identifies the Axillary Artery and is able to recognize and correct it	0.65	0.19 - 1.00
Q8S2A1: Exposes arteries by dissecting directly on anterior surface	0.32	0.09 - 0.52
Q8S2A2: Manipulates artery by grasping adventitia	0.36	0.13 - 0.56
Q8S2A3: Uses instruments properly	0.25	0.01 - 0.45
Q8S2A4: Positions body to use instruments to best advantage	0.32	0.08 - 0.53
Q8S2A5: Proceeds at appropriate pace	0.37	0.18 - 0.54
Q8S2A6: Handles tissue well with minimal damage	0.47	0.28 - 0.64
Q8S2A7: Creates an adequate visual field using retractors for the procedure	0.39	0.19 - 0.58
Q8S2A8: Communicates clearly and consistently	0.56	0.39 - 0.70
Q8S2A9: Performs the procedure without unnecessary dissection	0.65	0.51 - 0.76
Q8S2A10: Continually progresses towards the end goal	0.58	0.43 - 0.71
Q9S1A1: Operates through "key-hole" or too small a skin incision	0.36	0.16 - 0.55
Q9S1A2: Operates using full incision	0.07	-0.20 - 0.32
Q9S1A3: Excessive dissection	0.59	0.39 - 0.76
Q9S1A4: Pointless digging and shifting around in the surgical field	0.65	0.48 - 0.80
Q9S1A5: Has a logical operating sequence	0.51	0.31 - 0.71
Q9S1A6: Lacks anatomical knowledge	0.45	0.24 - 0.65
Q10S1A1: Improper instrument use	0.22	-0.02 - 0.44
Q10S1A2: Incorrect instrument holding	0.33	0.11 - 0.54
Q10S1A3: Scalpel use - multiple tentative cuts or cuts tangentially	0.50	0.28 - 0.70
Q10S1A4: Switches instruments excessively	0.51	0.29 - 0.70
Q10S1A5: Effective use of blunt dissection	0.21	-0.06 - 0.47
Q10S1A6: Dedicated use of a single instrument	0.06	-0.19 - 0.31
Q10S1A7: Uses sharp dissection confidently	0.34	0.00 - 0.62
Q11S1A1: Ligation of the axillary artery generally does not cause ischemia	0.55	0.34 - 0.73
Q12S1A1: Pitfall - bad incision	0.65	0.48 - 0.79
Q12S1A2: Pitfall - injures nerve, artery, or vein	0.48	0.17 - 0.74
Q12S1A3: Pitfall - inability to get proximal control	0.60	0.36 - 0.80
Q12S1A4: Pitfall - diving into clot/hematoma without adequate control	0.65	0.18 - 1.00
Q12S1A5: Pitfall - Mistaking nerve for artery	0.58	0.39 - 0.74

G1: Evaluation and treatment	0.52	0.38 - 0.64
G2: Anatomy	0.77	0.70 - 0.84
G3: Technical skill	0.45	0.25 - 0.62
G4: Overall readiness	0.77	0.67 - 0.85
Global score	0.80	0.68 - 0.89
IPS	0.84	0.78 - 0.89

APPENDIX D

DISCUSSION OF UNRELIABLE WSU AND MARYLAND VARIABLES

Among the WSU variables generated from transcripts of the procedures, *realizing a mistake* was found to be unreliable and was combined with *realizing that the surgeon is lost*. The resulting variable demonstrated 100% reliability in the recoded subsample. *Anticipating issues* and *facilitation* were not reliable. *Evaluating structures (appearance)*, *evaluating structures (process of elimination)*, and *evaluating structures (other cues)* also failed to meet the alpha threshold of 0.6. Subsequent attempts to improve reliability by combining these variables with other variables were either not conceptually justified or failed to improve reliability; these variables were not analyzed further.

Among the WSU video-based variables, *cuts per second during the opening incision*, *seconds per cut during the opening incision*, the *number of cuts to make the initial incision,* and the *total amount of time to make the incision* were unreliable and were therefore dropped from further analyses. The *proportion of the time using blunt dissection during the incision* and *proportion of the time using blunt dissection during the control phase* of the procedure likewise failed to meet the alpha threshold of 0.60. Each of these variables was dropped from further analyses as I did not have a conceptual justification for combining these variables with other variables to improve reliability.

As mentioned in the main body of the document, items within each main heading of the Maryland evaluation form were combined to form aggregated items. The resulting Maryland variables and alphas are listed in Table A4.1 below. Of the resulting 17 items, seven items had an alpha below 0.60. These items were dropped from subsequent analyses.

Table A4.1
Krippendorff's Alpha results for the combined items of the Maryland evaluation form.

Evaluation item	*Alpha*	*95% CI*
Q1: What structures may be injured?	0.69	0.53 - 0.80
Q2: What physical findings would you look for?	0.59	0.46 - 0.70
Q3: What additional studies would you use?	0.73	0.58 - 0.85
Q4: What is your plan for the patient?	0.39	-0.43 - 1.00
Q5: What is your resuscitation plan?	0.34	0.06 - 0.58
Q6 S1: How would you position the patient on the operating table?	0.14	-0.29 - 0.49
Q6 S2: How would you prep the patient for surgery?	0.56	0.41 - 0.68
Q7: What landmarks and incision would you use?	0.81	0.73 - 0.87
Q8 S1: Proper steps of the procedure	0.92	0.89 - 0.94
Q8 S2: Proper technique	0.63	0.50 - 0.75
Q9: Expert operative field maneuvers	0.66	0.47 - 0.80
Q10: Expert instrument use	0.25	-0.04 - 0.51
Q11: Consequences of ligating the artery	0.55	0.34 - 0.73
Q12: Common pitfalls of the procedure	0.69	0.56 - 0.80

APPENDIX E

REJECTED FACTOR MODEL CANDIDATES

11 factor model.

Rotated factor loadings for the model retaining 11 factors are listed in Table E.1 below. Factor 1 consisted of *Time between identifying the artery and placing the vessel loop, Total instrument changes, Instrument changes during the muscle phase, Instrument changes during the control phase,* and *Time spent searching for instruments.* Factor 2 consisted of *Extending the incision, Evaluating structures (by feel),* and *Backtracking.* Factor 3 consisted of *Weighing options, Evaluating progress, Risk mitigation,* and *Accounting for individual anatomy.* Factor 4 was composed of *Proportion of the time using sharp dissection during the muscle phase* and *Proportion of the time using sharp dissection during the identification phase.* Factor 5 included *Evaluating structures (location)* and *Evaluator assistance*. Factor 6 was made up of *Naming structures* and *Knowledge.* Factor 7 consisted of the *Proportion of the time using instruments in two hands during the muscle, identification, and control phases.* Factor 8 was composed of *Repositioning retractors.* Factor 9 included *Environment adjustment (self)* and *Shifts between blunt and sharp dissection*. Factor 10 was made up of *Proportion of the time using sharp dissection during the incision phase.* No variables met the loading criteria for Factor 11. The lack of any variables in Factor 11 and the presence of only one variable each in Factors 8 and 10 indicated that this model did not best capture the data; investigation of this model was therefore suspended.

Table E.1

Rotated factor loadings for the 11 factor model.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11
Forms a plan	0.55	0.03	0.52	0.09	0.02	0.18	0.05	0.05	0.06	0.15	0.13
Weighing options	-0.17	0.08	**0.66**	-0.06	0.09	0.28	-0.15	0.04	0.04	0.02	0.06
Expressions of confidence or certainty	-0.04	0.11	0.10	0.07	0.31	0.04	-0.22	0.29	0.35	-0.20	-0.33
Expressions of doubt or uncertainty	-0.08	0.64	0.17	-0.04	0.32	0.12	-0.10	0.12	0.36	-0.14	0.26
Mentioning things they expect to happen	0.00	-0.03	0.38	-0.16	0.34	0.49	0.25	-0.01	0.10	0.25	-0.18
Double checking behaviors	0.03	0.04	0.02	0.02	-0.06	-0.09	-0.13	0.07	-0.03	0.00	-0.69
Evaluating progress	0.21	0.01	**0.72**	0.02	0.24	0.27	-0.15	-0.19	0.10	0.13	-0.09
Risk mitigation	0.33	0.24	**0.64**	0.12	0.06	0.10	-0.04	0.07	-0.03	-0.17	-0.25
Accounting for individual anatomy	-0.17	-0.08	**0.64**	0.09	-0.08	-0.19	0.14	0.05	-0.16	0.15	0.00
Workarounds	0.11	0.28	0.42	0.14	0.22	0.02	0.15	0.43	-0.13	0.18	0.19
Naming structures	0.07	0.09	0.15	0.14	-0.03	**0.74**	-0.01	0.11	0.21	-0.01	-0.10
Evaluating structures (location)	0.10	0.17	0.05	0.05	**0.66**	0.00	-0.14	0.02	-0.12	-0.35	-0.06
Evaluating structures (behavior)	-0.11	0.10	0.17	0.21	0.22	0.07	0.01	0.51	-0.01	0.36	0.13
Knowledge	0.12	0.00	0.03	0.12	0.11	**0.64**	-0.12	-0.03	-0.01	-0.14	0.14
Heuristics	0.21	0.25	0.47	0.20	0.38	-0.26	0.01	0.26	0.02	-0.02	0.05
Balancing constraints	0.14	0.01	0.37	0.01	-0.36	0.47	-0.21	-0.04	-0.06	-0.02	0.24
Evaluator hint	0.00	0.51	0.06	-0.04	-0.08	0.08	-0.10	0.14	-0.02	0.25	0.45
Evaluator prompting	0.10	0.11	-0.16	-0.12	0.15	0.29	-0.17	0.15	0.54	0.32	0.15
Altering a vessel loop	-0.20	0.17	0.05	0.26	0.13	-0.20	0.09	0.38	-0.01	0.18	-0.16
Miscellaneous oddities	0.26	-0.10	0.20	0.35	-0.18	-0.12	0.03	0.46	0.11	0.05	0.10
Time between identifying the artery and placing the vessel loop	**0.72**	-0.09	-0.01	-0.09	-0.12	-0.14	-0.06	-0.23	0.05	-0.30	0.06
Environment adjustment (workspace)	0.42	0.11	0.35	-0.10	-0.13	-0.16	0.11	-0.03	0.13	-0.17	0.33
Environment adjustment (self)	-0.05	-0.04	-0.02	0.05	-0.03	-0.04	0.12	0.02	**0.84**	-0.03	0.03
Environment adjustment (patient)	0.28	-0.05	0.46	0.06	0.00	-0.07	0.03	0.34	-0.14	-0.23	0.20
Repositioning retractors	-0.04	0.04	-0.11	0.03	0.04	0.10	-0.02	**0.72**	0.18	-0.16	-0.13
Placing retractors or holding the incision open	0.03	0.06	-0.14	-0.24	0.60	0.21	-0.04	0.39	0.03	-0.03	0.01
Laying out instruments ahead of time	-0.12	-0.13	-0.06	-0.11	0.00	0.35	0.40	-0.05	-0.11	0.01	0.05
Extending the incision	0.12	**0.74**	0.11	0.13	0.06	-0.20	-0.03	-0.19	0.11	0.22	-0.06
Proportion of the time using instruments in two hands during the incision phase	0.12	0.24	0.03	-0.04	-0.24	0.08	0.55	-0.02	0.31	-0.13	-0.27
Proportion of the time using instruments in two hands during the muscle phase	0.25	0.06	-0.03	0.23	0.05	-0.02	**0.75**	0.04	0.22	0.00	-0.02
Proportion of the time using instruments in two hands during the identification phase	0.06	0.00	-0.08	0.15	-0.17	-0.09	**0.71**	0.15	0.07	0.02	0.17
Proportion of the time using instruments in two hands during the control phase	0.02	0.03	0.13	0.03	0.04	-0.38	**0.65**	-0.18	-0.10	-0.08	0.23
Total instrument changes	**0.84**	0.20	0.11	-0.11	0.18	0.15	0.20	0.08	-0.07	0.13	-0.06
Instrument changes during the incision phase	0.08	0.59	-0.10	-0.12	-0.11	0.20	0.51	-0.07	0.11	-0.22	-0.25
Instrument changes during the muscle phase	**0.73**	-0.14	-0.03	-0.07	0.24	-0.07	0.13	-0.16	0.02	0.35	0.04
Instrument changes during the identification phase	0.39	0.32	0.32	-0.10	0.12	0.24	0.02	0.36	-0.15	0.03	-0.08
Instrument changes during the control phase	**0.66**	-0.08	-0.08	0.11	-0.05	0.08	-0.02	0.04	-0.15	-0.27	0.05
Shifts between blunt and sharp dissection	-0.09	0.24	-0.03	0.26	-0.13	0.09	0.26	0.08	**0.68**	-0.20	-0.06
Proportion of the time sharp dissection used during the incision phase	-0.01	-0.02	0.09	0.11	-0.16	-0.12	-0.13	0.02	-0.12	**0.78**	0.04
Proportion of the time sharp dissection used during the muscle phase	0.11	0.07	0.07	**0.83**	-0.02	-0.13	0.01	0.29	0.10	0.07	-0.11
Proportion of the time sharp dissection used during the identification phase	-0.12	0.01	0.04	**0.74**	0.00	0.32	0.18	-0.14	0.01	-0.02	0.03
Proportion of the time sharp dissection used during the control phase	0.13	0.44	-0.13	0.21	-0.23	0.10	0.07	-0.46	-0.14	-0.15	0.14
Proportion of the time blunt dissection used during the muscle phase	-0.11	-0.07	-0.07	-0.83	0.02	0.13	-0.01	-0.29	-0.10	-0.07	0.11
Proportion of the time blunt dissection used during the identification phase	0.14	-0.03	0.02	-0.75	0.02	-0.36	-0.06	0.24	0.05	0.01	-0.12
Time spent searching for instruments	**0.77**	0.29	0.08	0.12	0.15	0.13	0.06	0.12	-0.01	0.12	-0.24
Idle time	0.50	0.54	0.19	0.14	0.15	-0.05	0.19	0.16	0.29	-0.08	0.03
Exploration	-0.03	0.58	-0.01	-0.02	0.16	0.10	0.11	0.08	0.51	0.04	-0.06
Evaluating structures (by feel)	-0.03	**0.74**	-0.04	-0.01	0.13	0.06	-0.03	0.02	-0.10	-0.22	-0.03
Checking by feel	0.30	0.23	0.34	-0.02	0.30	-0.33	-0.06	-0.03	0.00	0.18	0.39
Backtracking	0.06	**0.81**	0.13	0.09	0.14	-0.10	0.11	0.14	0.04	0.12	-0.02
Realizing a mistake	-0.02	0.28	0.36	0.09	0.58	-0.12	-0.18	0.21	0.13	-0.11	0.24
Evaluator assistance	0.30	0.16	0.26	-0.05	**0.68**	0.06	-0.04	-0.06	0.04	0.24	0.06

10 factor model.

Rotated factor loadings for the model retaining 10 factors are listed in Table E.2 below. Factor 1 was composed of *Extending the incision, Evaluating structures (by feel),* and *Backtracking.* Factor 2 consisted of *Time between identifying the artery and placing the vessel loop, Total instrument changes, Instrument changes during the muscle phase, Instrument changes during the control phase,* and *Time spent searching for instruments.* Factor 3 was made up of *Risk mitigation, Accounting for individual anatomy,* and *Environment adjustment (patient).* Factor 4 was composed of *Proportion of the time using sharp dissection during the muscle phase.* Factor 5 included *Evaluating structures (location)* and *Placing retractors or holding the incision open.* Factor 6 consisted of *Naming structures* and *Knowledge.* Factor 7 was composed of *Proportion of the time using instruments in two hands during the muscle, identification, and control phases.* Factor 8 was made up of *Environment adjustment (self)* and *Shifts between blunt and sharp dissection.* Factor 9 included *Proportion of the time using sharp dissection during the incision phase.* Finally, Factor 10 was composed of *Double checking behaviors*. As with the 11-factor model, three factors were composed of only one variable each. This model was rejected in favor of models containing factors with multiple variables.

Table E.2

Rotated factor loadings for the 10 factor model.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
Forms a plan	0.05	0.51	0.54	0.08	-0.02	0.25	0.04	0.07	0.15	-0.10
Weighing options	0.09	-0.21	0.60	-0.11	0.08	0.36	-0.16	0.02	0.07	-0.04
Expressions of confidence or certainty	0.11	-0.05	0.11	0.12	0.39	0.05	-0.20	0.34	-0.19	0.34
Expressions of doubt or uncertainty	0.65	-0.09	0.19	-0.04	0.32	0.12	-0.09	0.37	-0.14	-0.20
Mentioning things they expect to happen	0.00	0.04	0.27	-0.23	0.32	0.54	0.24	0.02	0.25	0.10
Double checking behaviors	0.03	0.04	-0.03	0.05	-0.03	-0.07	-0.11	-0.07	0.01	**0.66**
Evaluating progress	0.09	0.20	0.61	-0.08	0.08	0.42	-0.16	-0.03	0.11	-0.03
Risk mitigation	0.24	0.27	**0.63**	0.10	0.02	0.19	-0.03	-0.05	-0.15	0.27
Accounting for individual anatomy	-0.07	-0.21	**0.62**	0.09	-0.09	-0.09	0.13	-0.19	0.20	-0.01
Workarounds	0.24	0.07	0.51	0.22	0.35	0.02	0.14	-0.05	0.22	-0.06
Naming structures	0.06	0.06	0.09	0.10	0.04	**0.74**	0.00	0.24	0.00	0.17
Evaluating structures (location)	0.21	0.12	0.06	0.03	**0.57**	0.04	-0.15	-0.19	-0.41	-0.04
Evaluating structures (behavior)	0.06	-0.11	0.24	0.31	0.41	0.04	0.01	0.08	0.40	-0.01
Knowledge	-0.01	0.12	0.00	0.06	0.09	**0.64**	-0.12	0.02	-0.16	-0.11
Heuristics	0.28	0.17	0.54	0.25	0.38	-0.19	0.01	0.00	-0.02	-0.05
Balancing constraints	-0.03	0.08	0.36	-0.04	-0.34	0.48	-0.21	0.03	0.02	-0.11
Evaluator hint	0.48	-0.02	0.12	-0.01	0.00	0.04	-0.10	0.09	0.27	-0.29
Evaluator prompting	0.12	0.14	-0.17	-0.09	0.23	0.24	-0.16	0.56	0.31	-0.11
Altering a vessel loop	0.14	-0.20	0.09	0.35	0.26	-0.20	0.10	0.02	0.21	0.21
Miscellaneous oddities	-0.16	0.19	0.33	0.47	-0.01	-0.14	0.04	0.23	0.11	0.06
Time between identifying the artery and placing the vessel loop	-0.07	**0.70**	0.06	-0.11	-0.26	-0.13	-0.06	0.04	-0.35	-0.09
Environment adjustment (workspace)	0.11	0.36	0.44	-0.09	-0.18	-0.14	0.11	0.17	-0.15	-0.27
Environment adjustment (self)	0.00	-0.05	-0.05	0.05	-0.03	-0.01	0.15	**0.79**	-0.02	-0.09
Environment adjustment (patient)	-0.10	0.20	**0.59**	0.14	0.09	-0.08	0.02	-0.02	-0.18	-0.04
Repositioning retractors	-0.07	-0.08	0.04	0.20	0.38	-0.05	0.00	0.38	-0.08	0.37
Placing retractors or holding the incision open	0.03	0.07	-0.07	-0.16	**0.74**	0.11	-0.05	0.09	-0.04	0.05
Laying out instruments ahead of time	-0.15	-0.10	-0.10	-0.15	0.02	0.32	0.39	-0.10	0.01	-0.04
Extending the incision	**0.79**	0.13	0.06	0.10	-0.06	-0.11	-0.02	0.01	0.17	-0.01
Proportion of the time using instruments in two hands during the incision phase	0.22	0.10	0.02	-0.05	-0.21	0.08	**0.58**	0.29	-0.12	0.28
Proportion of the time using instruments in two hands during the muscle phase	0.06	0.25	-0.01	0.23	0.04	0.00	**0.75**	0.18	-0.02	-0.03
Proportion of the time using instruments in two hands during the identification phase	-0.04	0.04	-0.01	0.19	-0.08	-0.13	**0.72**	0.12	0.05	-0.11
Proportion of the time using instruments in two hands during the control phase	0.06	0.01	0.15	0.00	-0.07	-0.32	**0.64**	-0.17	-0.09	-0.32
Total instrument changes	0.18	**0.85**	0.18	-0.07	0.17	0.12	0.20	-0.05	0.08	0.09
Instrument changes during the incision phase	0.56	0.07	-0.11	-0.15	-0.10	0.16	0.54	0.10	-0.22	0.29
Instrument changes during the muscle phase	-0.08	**0.78**	-0.03	-0.08	0.11	-0.03	0.11	-0.06	0.26	-0.17
Instrument changes during the identification phase	0.25	0.36	0.41	-0.03	0.25	0.19	0.02	-0.04	0.05	0.24
Instrument changes during the control phase	-0.11	**0.63**	0.02	0.13	-0.06	0.05	-0.03	-0.08	-0.30	0.00
Shifts between blunt and sharp dissection	0.25	-0.12	-0.05	0.25	-0.10	0.12	0.30	**0.65**	-0.18	0.07
Proportion of the time sharp dissection used during the incision phase	0.00	0.02	0.04	0.14	-0.14	-0.09	-0.14	-0.13	**0.78**	-0.05
Proportion of the time sharp dissection used during the muscle phase	0.07	0.07	0.11	**0.89**	0.03	-0.05	0.03	0.10	0.07	0.11
Proportion of the time sharp dissection used during the identification phase	0.05	-0.13	-0.05	0.64	-0.11	0.44	0.19	-0.08	-0.06	-0.13
Proportion of the time sharp dissection used during the control phase	0.46	0.13	-0.18	0.10	-0.42	0.16	0.08	-0.21	-0.21	-0.19
Proportion of the time blunt dissection used during the muscle phase	-0.07	-0.07	-0.11	-0.89	-0.03	0.05	-0.03	-0.10	-0.07	-0.11
Proportion of the time blunt dissection used during the identification phase	-0.07	0.14	0.13	-0.63	0.17	-0.49	-0.07	0.15	0.06	0.23
Time spent searching for instruments	0.28	**0.77**	0.13	0.15	0.14	0.13	0.07	-0.01	0.07	0.26
Idle time	0.54	0.47	0.27	0.18	0.15	-0.03	0.20	0.30	-0.10	0.01
Exploration	0.60	-0.02	-0.03	-0.02	0.18	0.11	0.14	0.48	0.03	0.07
Evaluating structures (by feel)	**0.71**	-0.04	-0.01	-0.02	0.12	0.04	-0.02	-0.08	-0.24	0.10
Checking by feel	0.28	0.29	0.39	-0.01	0.19	-0.26	-0.08	-0.04	0.16	-0.43
Evaluator assistance	0.24	0.35	0.22	-0.09	0.55	0.14	-0.06	-0.09	0.17	-0.20
Backtracking	**0.80**	0.05	0.16	0.13	0.17	-0.08	0.12	0.04	0.12	0.08
Realizing a mistake	0.32	-0.03	0.41	0.12	0.56	-0.06	-0.19	0.10	-0.12	-0.25

9 factor model.

Rotated factor loadings for the model containing nine factors are listed in Table E.3 below. Factor 1 consisted of *Extending the incision, Backtracking, Evaluating structures (by feel),* and *Expressions of doubt or uncertainty.* Factor 2 was composed of *Time between identifying the artery and placing the vessel loop, Total instrument changes, Instrument changes during the muscle phase, Instrument changes during the control phase,* and *Time spent searching for instruments.* Factor 3 was made up of *Accounting for individual anatomy* and *Environment adjustment (patient).* Factor 4 was composed of *Proportion of the time using sharp dissection during the muscle phase.* Factor 5 included *Naming structures, Knowledge,* and *Balancing constraints.* Factor 6 included *Proportion of the time using instruments in two hands during the incision, muscle, and identification phases.* Factor 7 was composed of *Placing retractors or holding the incision open, Expressions of confidence or certainty,* and *Repositioning retractors.* Factor 8 was made up of *Proportion of the time using sharp dissection during the incision phase.* Factor 9 included *Environment adjustment (self)* and *Evaluator prompting*. As before, multiple factors only contained one variable; this model was similarly rejected.

Table E.3

Rotated factor loadings for the nine factor model.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9
Double checking behaviors	-0.03	0.05	-0.14	0.12	-0.02	0.02	0.38	-0.07	-0.27
Accounting for individual anatomy	-0.10	-0.20	**0.58**	0.09	0.00	0.11	-0.16	0.20	-0.20
Naming structures	0.06	0.06	-0.04	0.15	**0.75**	0.11	0.16	0.01	0.16
Knowledge	0.03	0.11	-0.05	0.06	**0.62**	-0.11	0.04	-0.09	0.04
Time between identifying the artery and placing the vessel loop	-0.09	**0.63**	0.08	-0.08	-0.09	-0.04	-0.16	-0.52	0.06
Environment adjustment (self)	0.01	-0.08	-0.03	0.07	0.00	0.22	0.10	-0.08	**0.76**
Environment adjustment (patient)	-0.09	0.18	**0.63**	0.10	0.01	0.00	0.08	-0.13	-0.03
Placing retractors or holding the incision open	0.12	0.13	0.01	-0.28	0.09	-0.10	**0.57**	0.25	0.08
Extending the incision	**0.76**	0.14	0.04	0.13	-0.09	0.03	-0.14	0.10	0.01
Proportion of the time using instruments in two hands during the muscle phase	0.05	0.25	0.05	0.21	-0.08	**0.73**	0.00	0.05	0.15
Proportion of the time using instruments in two hands during the identification phase	-0.06	0.04	0.06	0.18	-0.20	**0.67**	-0.15	0.08	0.12
Proportion of the time using instruments in two hands during the control phase	0.06	0.00	0.28	-0.06	-0.37	0.53	-0.30	-0.02	-0.10
Total instrument changes	0.17	**0.87**	0.17	-0.08	0.14	0.20	0.07	0.05	-0.07
Instrument changes during the muscle phase	-0.09	**0.80**	0.02	-0.11	-0.06	0.04	-0.14	0.20	0.04
Shifts between blunt and sharp dissection	0.23	-0.15	-0.06	0.30	0.10	0.41	0.14	-0.21	0.56
Proportion of the time sharp dissection used during the muscle phase	0.07	0.08	0.14	**0.88**	-0.08	0.03	0.14	0.10	0.05
Time spent searching for instruments	0.26	**0.79**	0.09	0.17	0.15	0.12	0.19	0.02	-0.08
Backtracking	**0.79**	0.08	0.17	0.12	-0.06	0.16	0.09	0.16	-0.01
Evaluating structures (by feel)	**0.73**	-0.04	-0.01	-0.02	0.05	0.03	0.15	-0.16	-0.14
Forms a plan	0.03	0.51	0.49	0.09	0.33	0.06	-0.13	0.08	0.09
Weighing options	0.10	-0.21	0.51	-0.10	0.47	-0.12	0.01	0.11	0.01
Expressions of confidence or certainty	0.15	-0.04	0.11	0.09	0.10	-0.12	**0.61**	-0.07	0.20
Expressions of doubt or uncertainty	**0.71**	-0.09	0.23	-0.09	0.16	-0.06	0.18	-0.01	0.38
Mentioning things they expect to happen	0.00	0.09	0.18	-0.24	0.56	0.26	0.18	0.37	-0.02
Evaluating progress	0.08	0.21	0.51	-0.07	0.53	-0.13	-0.03	0.11	-0.03
Risk mitigation	0.22	0.25	0.55	0.13	0.32	0.06	0.17	-0.18	-0.18
Workarounds	0.26	0.12	0.55	0.15	0.07	0.10	0.12	0.36	-0.04
Evaluating structures (location)	0.31	0.13	0.18	-0.09	0.03	-0.21	0.43	-0.13	-0.19
Evaluating structures (behavior)	0.09	-0.04	0.26	0.24	0.05	-0.04	0.22	0.55	0.09
Heuristics	0.32	0.20	0.62	0.17	-0.11	-0.04	0.23	0.13	-0.01
Balancing constraints	-0.07	0.04	0.22	0.04	**0.56**	-0.14	-0.32	-0.13	0.05
Evaluator hint	0.48	0.01	0.12	-0.01	0.05	-0.12	-0.25	0.25	0.18
Evaluator prompting	0.15	0.17	-0.18	-0.09	0.23	-0.12	0.13	0.30	**0.60**
Altering a vessel loop	0.15	-0.15	0.12	0.31	-0.21	0.09	0.29	0.32	-0.05
Miscellaneous oddities	-0.17	0.19	0.35	0.46	-0.09	0.05	0.08	0.07	0.20
Environment adjustment (workspace)	0.10	0.31	0.47	-0.09	-0.07	0.11	-0.26	-0.25	0.22
Repositioning retractors	-0.04	-0.06	0.05	0.17	-0.03	0.07	**0.62**	0.04	0.23
Laying out instruments ahead of time	-0.15	-0.09	-0.12	-0.16	0.26	0.36	-0.06	0.09	-0.10
Proportion of the time using instruments in two hands during the incision phase	0.15	0.08	-0.05	0.03	0.08	**0.69**	0.05	-0.20	0.14
Instrument changes during the incision phase	0.50	0.06	-0.17	-0.09	0.13	0.65	0.10	-0.24	-0.05
Instrument changes during the identification phase	0.25	0.38	0.35	-0.03	0.26	0.07	0.26	0.09	-0.13
Instrument changes during the control phase	-0.11	**0.59**	0.05	0.14	0.05	-0.03	0.01	-0.36	-0.08
Proportion of the time sharp dissection used during the incision phase	-0.06	0.09	-0.02	0.17	-0.08	-0.17	-0.33	**0.63**	-0.05
Proportion of the time sharp dissection used during the identification phase	0.05	-0.13	-0.06	0.64	0.35	0.16	-0.13	0.03	-0.06
Proportion of the time sharp dissection used during the control phase	0.43	0.08	-0.20	0.16	0.11	0.09	-0.44	-0.32	-0.16
Proportion of the time blunt dissection used during the muscle phase	-0.07	-0.08	-0.14	-0.88	0.08	-0.03	-0.14	-0.10	-0.05
Proportion of the time blunt dissection used during the identification phase	-0.09	0.15	0.14	-0.64	-0.39	-0.03	0.24	0.00	0.09
Idle time	0.55	0.46	0.30	0.17	0.01	0.25	0.14	-0.09	0.25
Exploration	0.60	0.00	-0.04	0.00	0.11	0.22	0.21	0.06	0.41
Checking by feel	0.32	0.31	0.50	-0.09	-0.21	-0.18	-0.20	0.20	0.10
Evaluator assistance	0.31	0.41	0.28	-0.19	0.15	-0.15	0.17	0.35	-0.01
Realizing a mistake	0.42	-0.01	0.53	-0.01	-0.01	-0.26	0.29	0.13	0.15

8 factor model.

Rotated factor loadings for the 8 factor model are listed in Table E.4 below. Factor 1 consisted of *Extending the incision, Backtracking, Evaluating structures (by feel),* and *Expressions of doubt or uncertainty.* Factor 2 was made up of *Time between identifying the artery and placing the vessel loop, Total instrument changes, Instrument changes during the muscle phase, Instrument changes during the control phase,* and *Time spent searching for instruments.* Factor 3 was made up of *Accounting for individual anatomy* and *Environment adjustment (patient).* Factor 4 was composed of *Proportion of the time using sharp dissection during the muscle phase.* Factor 5 included *Proportion of the time using instruments in two hands during the incision, muscle, and identification phases,* as well as *Shifts between blunt and sharp dissection.* Factor 6 included *Repositioning retractors* and *Placing retractors or holding the incision open.* Factor 7 was made up of *Naming structures, Knowledge,* and *Balancing constraints.* Finally, Factor 8 was composed of *Proportion of the time using sharp dissection during the incision phase.* As with the earlier models, the 8 factor model was rejected due to multiple factors containing only one variable.

Table E.4

Rotated factor loadings for the eight factor model.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8
Forms a plan	0.04	0.52	0.45	0.10	0.07	0.01	0.34	0.19
Weighing options	0.12	-0.20	0.52	-0.09	-0.12	0.09	0.46	0.05
Expressions of confidence or certainty	0.18	-0.04	0.06	0.15	-0.04	0.51	0.09	-0.38
Expressions of doubt or uncertainty	**0.75**	-0.08	0.15	-0.07	0.08	0.27	0.17	-0.04
Mentioning things they expect to happen	0.01	0.05	0.27	-0.20	0.19	0.32	0.51	0.17
Double checking behaviors	-0.05	0.05	-0.08	0.13	-0.08	0.12	-0.06	-0.34
Evaluating progress	0.10	0.22	0.51	-0.06	-0.15	0.07	0.53	0.09
Risk mitigation	0.22	0.29	0.55	0.12	0.00	-0.02	0.30	-0.28
Accounting for individual anatomy	-0.11	-0.19	**0.64**	0.10	0.02	-0.14	-0.02	0.17
Workarounds	0.28	0.09	0.58	0.22	0.04	0.20	0.05	0.18
Naming structures	0.07	0.05	-0.05	0.13	0.16	0.18	**0.75**	-0.08
Evaluating structures (location)	0.31	0.15	0.18	-0.05	-0.27	0.22	0.00	-0.38
Evaluating structures (behavior)	0.12	-0.11	0.29	0.34	-0.05	0.41	0.04	0.30
Knowledge	0.03	0.11	-0.06	0.02	-0.07	0.03	**0.63**	-0.09
Heuristics	0.34	0.21	0.60	0.24	-0.06	0.21	-0.12	-0.04
Balancing constraints	-0.06	0.07	0.18	-0.02	-0.09	-0.24	**0.59**	0.08
Evaluator hint	0.50	-0.02	0.09	0.00	-0.06	-0.02	0.07	0.37
Evaluator prompting	0.21	0.11	-0.27	-0.03	0.07	0.52	0.26	0.31
Altering a vessel loop	0.15	-0.19	0.17	0.38	0.04	0.25	-0.22	0.03
Miscellaneous oddities	-0.14	0.19	0.29	0.50	0.12	0.13	-0.06	0.05
Time between identifying the artery and placing the vessel loop	-0.09	**0.70**	-0.02	-0.14	0.01	-0.25	-0.07	-0.22
Environment adjustment (workspace)	0.12	0.37	0.38	-0.12	0.19	-0.18	-0.04	0.04
Environment adjustment (self)	0.06	-0.07	-0.17	0.08	0.48	0.34	0.04	0.05
Environment adjustment (patient)	-0.07	0.22	**0.59**	0.11	-0.02	0.01	0.01	-0.15
Repositioning retractors	-0.01	-0.07	0.03	0.24	0.13	**0.56**	-0.04	-0.30
Placing retractors or holding the incision open	0.15	0.08	0.03	-0.18	-0.11	**0.63**	0.05	-0.10
Laying out instruments ahead of time	-0.18	-0.10	-0.03	-0.18	0.29	-0.06	0.23	0.07
Extending the incision	**0.75**	0.13	0.03	0.13	0.03	-0.12	-0.09	0.15
Proportion of the time using instruments in two hands during the incision phase	0.13	0.11	-0.03	-0.02	**0.71**	-0.07	0.06	-0.17
Proportion of the time using instruments in two hands during the muscle phase	0.03	0.25	0.08	0.20	**0.73**	-0.01	-0.09	0.06
Proportion of the time using instruments in two hands during the identification phase	-0.08	0.05	0.10	0.16	**0.66**	-0.11	-0.20	0.16
Proportion of the time using instruments in two hands during the control phase	0.03	0.03	0.32	-0.08	0.46	-0.33	-0.38	0.13
Total instrument changes	0.16	**0.86**	0.18	-0.06	0.14	0.09	0.12	0.05
Instrument changes during the incision phase	0.46	0.09	-0.11	-0.15	0.61	-0.13	0.09	-0.28
Instrument changes during the muscle phase	-0.08	**0.77**	0.00	-0.07	0.01	0.07	-0.06	0.33
Instrument changes during the identification phase	0.25	0.38	0.38	0.01	-0.01	0.21	0.23	-0.09
Instrument changes during the control phase	-0.12	**0.63**	0.00	0.10	-0.03	-0.15	0.06	-0.25
Shifts between blunt and sharp dissection	0.25	-0.13	-0.15	0.27	**0.61**	0.16	0.14	-0.16
Proportion of the time sharp dissection used during the incision phase	-0.05	0.01	0.02	0.23	-0.21	-0.01	-0.06	**0.68**
Proportion of the time sharp dissection used during the muscle phase	0.08	0.07	0.11	**0.90**	0.07	0.04	-0.04	-0.04
Proportion of the time sharp dissection used during the identification phase	0.03	-0.14	-0.04	0.59	0.16	-0.24	0.38	0.02
Proportion of the time sharp dissection used during the control phase	0.39	0.12	-0.21	0.05	0.08	-0.60	0.13	-0.04
Proportion of the time blunt dissection used during the muscle phase	-0.08	-0.07	-0.11	-0.90	-0.07	-0.04	0.04	0.04
Proportion of the time blunt dissection used during the identification phase	-0.06	0.16	0.12	-0.57	-0.04	0.33	-0.43	-0.06
Time spent searching for instruments	0.25	**0.78**	0.09	0.19	0.07	0.13	0.14	-0.07
Idle time	0.56	0.48	0.23	0.18	0.33	0.14	0.02	-0.08
Exploration	0.63	-0.02	-0.09	0.02	0.35	0.31	0.11	0.00
Evaluating structures (by feel)	**0.71**	-0.02	0.01	-0.04	0.00	-0.07	0.03	-0.27
Checking by feel	0.35	0.31	0.44	-0.04	-0.16	0.01	-0.20	0.33
Evaluator assistance	0.33	0.37	0.29	-0.11	-0.19	0.33	0.13	0.21
Backtracking	**0.79**	0.07	0.18	0.15	0.14	0.06	-0.08	0.05
Realizing a mistake	0.46	-0.01	0.47	0.06	-0.21	0.35	-0.01	-0.04

7 factor model.

Rotated factor loadings for the model containing seven factors can be found in Table E.5 below. Factor 1 was made up of *Time between identifying the artery and placing the vessel loop, Total instrument changes, Instrument changes during the muscle phase, Instrument changes during the control phase,* and *Time spent searching for instruments.* Factor 2 consisted of *Extending the incision, Backtracking, Evaluating structures (by feel),* and *Expressions of doubt or uncertainty.* Factor 3 was composed of *Proportion of the time using sharp dissection during the muscle phase.* Factor 4 consisted of *Accounting for individual anatomy.* Factor 5 included *Proportion of the time using instruments in two hands during the incision, muscle, and identification phases,* as well as *Shifts between blunt and sharp dissection.* Factor 6 included *Expressions of confidence or certainty, Repositioning retractors,* and *Placing retractors or holding the incision open.* Factor 7 was made up of *Naming structures, Knowledge,* and *Balancing constraints.* This model also contained multiple factors composed of only one variable, so I did not consider it further.

Table E.5

Rotated factor loadings for the seven factor model.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Forms a plan	0.61	0.05	0.08	0.33	0.07	-0.02	0.37
Weighing options	-0.03	0.16	-0.07	0.41	-0.14	0.10	0.52
Expressions of confidence or certainty	0.01	0.18	0.18	-0.12	-0.07	**0.60**	0.08
Expressions of doubt or uncertainty	-0.02	**0.75**	-0.07	0.08	0.08	0.31	0.18
Mentioning things they expect to happen	0.12	0.00	-0.23	0.26	0.21	0.28	0.54
Double checking behaviors	0.04	-0.04	0.18	-0.24	-0.11	0.19	-0.08
Evaluating progress	0.35	0.13	-0.05	0.35	-0.16	0.07	0.57
Risk mitigation	0.47	0.27	0.20	0.17	-0.07	0.06	0.34
Accounting for individual anatomy	0.00	-0.04	0.11	**0.62**	-0.01	-0.16	0.06
Workarounds	0.25	0.30	0.19	0.56	0.04	0.19	0.10
Naming structures	0.04	0.04	0.15	-0.14	0.18	0.21	**0.73**
Evaluating structures (location)	0.22	0.33	0.01	-0.10	-0.32	0.31	0.01
Evaluating structures (behavior)	-0.05	0.10	0.26	0.49	0.00	0.36	0.06
Knowledge	0.09	0.02	0.05	-0.17	-0.07	0.06	**0.61**
Heuristics	0.38	0.38	0.24	0.44	-0.09	0.24	-0.08
Balancing constraints	0.12	-0.05	0.00	0.10	-0.10	-0.24	**0.60**
Evaluator hint	-0.03	0.48	-0.06	0.26	0.00	-0.08	0.09
Evaluator prompting	-0.01	0.12	-0.13	-0.02	0.17	0.44	0.22
Altering a vessel loop	-0.14	0.16	0.35	0.26	0.05	0.26	-0.21
Miscellaneous oddities	0.26	-0.14	0.47	0.28	0.12	0.13	-0.05
Time between identifying the artery and placing the vessel loop	**0.68**	-0.08	-0.09	-0.31	-0.03	-0.22	-0.09
Environment adjustment (workspace)	0.47	0.15	-0.10	0.20	0.15	-0.18	-0.01
Environment adjustment (self)	-0.11	0.03	0.04	-0.06	0.51	0.32	0.02
Environment adjustment (patient)	0.41	-0.01	0.16	0.33	-0.08	0.05	0.06
Repositioning retractors	-0.04	-0.02	0.25	-0.06	0.12	**0.61**	-0.05
Placing retractors or holding the incision open	0.09	0.12	-0.20	-0.01	-0.09	**0.64**	0.04
Laying out instruments ahead of time	-0.10	-0.18	-0.17	-0.01	0.29	-0.08	0.23
Extending the incision	0.12	**0.74**	0.10	0.06	0.06	-0.12	-0.09
Proportion of the time using instruments in two hands during the incision phase	0.13	0.14	0.02	-0.19	**0.68**	-0.04	0.05
Proportion of the time using instruments in two hands during the muscle phase	0.28	0.03	0.19	0.05	**0.73**	-0.03	-0.09
Proportion of the time using instruments in two hands during the identification phase	0.08	-0.07	0.14	0.16	**0.67**	-0.15	-0.20
Proportion of the time using instruments in two hands during the control phase	0.13	0.08	-0.08	0.29	0.42	-0.36	-0.34
Total instrument changes	**0.87**	0.13	-0.08	-0.01	0.14	0.08	0.11
Instrument changes during the incision phase	0.09	0.48	-0.08	-0.34	0.57	-0.07	0.08
Instrument changes during the muscle phase	**0.70**	-0.14	-0.15	0.07	0.07	-0.02	-0.08
Instrument changes during the identification phase	0.49	0.26	0.02	0.16	-0.03	0.24	0.25
Instrument changes during the control phase	**0.62**	-0.12	0.14	-0.27	-0.07	-0.10	0.03
Shifts between blunt and sharp dissection	-0.14	0.25	0.29	-0.19	**0.60**	0.20	0.11
Proportion of the time sharp dissection used during the incision phase	-0.06	-0.10	0.09	0.45	-0.11	-0.14	-0.05
Proportion of the time sharp dissection used during the muscle phase	0.09	0.07	**0.89**	0.13	0.08	0.08	-0.05
Proportion of the time sharp dissection used during the identification phase	-0.15	0.03	0.61	0.00	0.17	-0.21	0.36
Proportion of the time sharp dissection used during the control phase	0.06	0.39	0.09	-0.30	0.07	-0.57	0.11
Proportion of the time blunt dissection used during the muscle phase	-0.09	-0.07	-0.89	-0.13	-0.08	-0.08	0.05
Proportion of the time blunt dissection used during the identification phase	0.20	-0.06	-0.58	0.06	-0.06	0.31	-0.41
Time spent searching for instruments	**0.77**	0.22	0.18	-0.11	0.08	0.14	0.11
Idle time	0.54	0.56	0.18	0.03	0.31	0.17	0.02
Exploration	-0.04	0.60	-0.01	-0.08	0.38	0.32	0.10
Evaluating structures (by feel)	0.01	**0.72**	0.01	-0.19	-0.04	0.02	0.02
Checking by feel	0.40	0.35	-0.10	0.48	-0.14	-0.04	-0.15
Evaluator assistance	0.41	0.31	-0.17	0.29	-0.15	0.29	0.15
Backtracking	0.12	**0.79**	0.13	0.15	0.15	0.08	-0.06
Realizing a mistake	0.13	0.49	0.06	0.37	-0.22	0.39	0.03

6 factor model.

Rotated factor loadings for the six factor model are listed in Table E.6 below. Factor 1 included *Expressions of doubt or uncertainty, Evaluator hints, Extending the incision, Evaluating structures (by feel), Backtracking,* and *Realizing a mistake*. Factor 2 was made up of *Time between identifying the artery and placing the vessel loop, Total instrument changes, Instrument changes during the muscle phase, Instrument changes during the control phase,* and *Time spent searching for instruments.* Factor 3 was composed of *Proportion of the time using instruments in two hands during the incision, Instrument changes during the incision,* and *Shifts between blunt and sharp dissection.* Factor 4 consisted of *Proportion of the time using sharp dissection during the muscle phase.* Factor 5 was made up of *Naming structures, Knowledge,* and *Balancing constraints.* Finally, Factor 6 contained *Expressions of confidence or certainty, Repositioning retractors,* and *Placing retractors or holding the incision open.* Investigation of this model was suspended due to the presence of only one variable in Factor 4.

Table E.6

Rotated factor loadings for the six factor model.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Forms a plan	0.16	0.62	-0.09	0.23	0.38	-0.08
Weighing options	0.31	-0.03	-0.30	0.09	0.53	0.03
Expressions of confidence or certainty	0.16	-0.03	0.11	0.13	0.11	**0.60**
Expressions of doubt or uncertainty	**0.76**	-0.06	0.20	-0.04	0.18	0.24
Mentioning things they expect to happen	0.14	0.15	-0.01	-0.03	0.51	0.20
Double checking behaviors	-0.12	0.01	0.07	0.05	-0.05	0.24
Evaluating progress	0.26	0.34	-0.28	0.08	0.59	0.02
Risk mitigation	0.30	0.44	-0.03	0.22	0.37	0.04
Accounting for individual anatomy	0.15	0.03	-0.34	0.37	0.07	-0.26
Workarounds	0.47	0.24	-0.16	0.43	0.12	0.08
Naming structures	-0.01	0.03	0.28	0.07	**0.73**	0.19
Evaluating structures (location)	0.31	0.17	-0.10	-0.08	0.04	0.35
Evaluating structures (behavior)	0.27	-0.05	-0.19	0.48	0.08	0.26
Knowledge	-0.04	0.07	0.08	-0.07	**0.62**	0.09
Heuristics	0.51	0.35	-0.17	0.41	-0.04	0.17
Balancing constraints	-0.03	0.12	-0.12	-0.01	**0.61**	-0.25
Evaluator hint	**0.53**	-0.03	-0.04	0.02	0.09	-0.14
Evaluator prompting	0.16	0.00	0.15	-0.08	0.21	0.40
Altering a vessel loop	0.21	-0.16	0.00	0.46	-0.19	0.19
Miscellaneous oddities	-0.07	0.26	0.00	0.59	-0.02	0.08
Time between identifying the artery and placing the vessel loop	-0.18	**0.67**	0.11	-0.22	-0.08	-0.13
Environment adjustment (workspace)	0.22	0.49	0.02	0.01	-0.02	-0.22
Environment adjustment (self)	0.02	-0.09	0.45	0.11	-0.02	0.26
Environment adjustment (patient)	0.10	0.40	-0.21	0.29	0.08	0.01
Repositioning retractors	-0.01	-0.06	0.17	0.27	-0.04	**0.59**
Placing retractors or holding the incision open	0.20	0.08	-0.07	-0.13	0.05	**0.64**
Laying out instruments ahead of time	-0.16	-0.06	0.16	-0.12	0.19	-0.11
Extending the incision	**0.68**	0.09	0.21	0.06	-0.07	-0.16
Proportion of the time using instruments in two hands during the incision phase	0.05	0.16	**0.67**	0.02	0.00	-0.08
Proportion of the time using instruments in two hands during the muscle phase	0.01	0.31	0.57	0.30	-0.13	-0.11
Proportion of the time using instruments in two hands during the identification phase	-0.05	0.13	0.42	0.30	-0.24	-0.25
Proportion of the time using instruments in two hands during the control phase	0.15	0.19	0.15	0.11	-0.38	-0.44
Total instrument changes	0.16	**0.87**	0.14	-0.04	0.11	0.08
Instrument changes during the incision phase	0.32	0.10	**0.73**	-0.20	0.04	-0.09
Instrument changes during the muscle phase	-0.07	**0.72**	-0.05	-0.06	-0.08	-0.01
Instrument changes during the identification phase	0.33	0.47	-0.03	0.09	0.27	0.21
Instrument changes during the control phase	-0.22	**0.59**	0.10	-0.01	0.05	-0.02
Shifts between blunt and sharp dissection	0.13	-0.14	**0.69**	0.24	0.09	0.14
Proportion of the time sharp dissection used during the incision phase	0.04	-0.04	-0.35	0.28	-0.04	-0.20
Proportion of the time sharp dissection used during the muscle phase	0.01	0.04	0.18	**0.84**	0.01	0.04
Proportion of the time sharp dissection used during the identification phase	-0.07	-0.18	0.28	0.50	0.39	-0.24
Proportion of the time sharp dissection used during the control phase	0.20	0.04	0.34	-0.15	0.11	-0.53
Proportion of the time blunt dissection used during the muscle phase	-0.01	-0.04	-0.18	-0.84	-0.01	-0.04
Proportion of the time blunt dissection used during the identification phase	0.08	0.23	-0.23	-0.42	-0.44	0.32
Time spent searching for instruments	0.17	**0.74**	0.22	0.12	0.14	0.16
Idle time	0.53	0.52	0.40	0.20	0.02	0.12
Exploration	0.55	-0.05	0.48	0.00	0.08	0.26
Evaluating structures (by feel)	**0.60**	-0.04	0.26	-0.14	0.04	0.02
Checking by feel	0.51	0.40	-0.31	0.11	-0.14	-0.10
Evaluator assistance	0.45	0.40	-0.23	-0.02	0.16	0.26
Backtracking	**0.77**	0.08	0.25	0.16	-0.05	0.01
Realizing a mistake	**0.61**	0.10	-0.24	0.20	0.06	0.33

4 factor model.

Rotated factor loadings for the model containing four factors are presented in Table E.7 below. Factor 1 contained *Expressions of doubt or uncertainty, Placing retractors or holding the incision open,* and *Realizing a mistake.* Factor 2 was made up of *Forms a plan, Time between identifying the artery and placing the vessel loop, Environment adjustment (workspace), Total instrument changes, Instrument changes during the muscle phase, Instrument changes during the control phase,* and *Time spent searching for instruments.* Factor 3 included *Proportion of the time using instruments in two hands during the incision and muscle phases, Instrument changes during the incision,* and *Shifts between blunt and sharp dissection.* Factor 4 contained *Miscellaneous oddities* and *Proportion of the time using sharp dissection during the muscle and identification phases.* Although each of the four factors in this model contained multiple variables, Factors 1, 2, and 4 lacked obvious coherence and did not appear to capture any higher order constructs as effectively as the model retaining five factors. The model containing four factors was therefore rejected.

Table E.7

Rotated factor loadings for the four factor model.

Variable	Factor 1	Factor 2	Factor 3	Factor 4
Forms a plan	0.24	**0.68**	-0.05	0.23
Weighing options	0.43	0.08	-0.31	0.12
Expressions of confidence or certainty	0.48	-0.17	0.04	0.08
Expressions of doubt or uncertainty	**0.77**	-0.08	0.28	-0.03
Mentioning things they expect to happen	0.38	0.16	-0.07	-0.02
Double checking behaviors	0.02	-0.07	0.03	0.02
Evaluating progress	0.42	0.44	-0.27	0.10
Risk mitigation	0.40	0.48	0.01	0.22
Accounting for individual anatomy	0.03	0.16	-0.29	0.38
Workarounds	0.48	0.28	-0.08	0.41
Naming structures	0.30	0.02	0.13	0.13
Evaluating structures (location)	0.46	0.11	-0.06	-0.13
Evaluating structures (behavior)	0.40	-0.06	-0.20	0.45
Knowledge	0.19	0.10	-0.03	-0.03
Heuristics	0.52	0.34	-0.06	0.35
Balancing constraints	0.03	0.25	-0.17	0.06
Evaluator hint	0.37	0.04	0.06	0.06
Evaluator prompting	0.39	-0.09	0.10	-0.10
Altering a vessel loop	0.22	-0.20	0.01	0.42
Miscellaneous oddities	0.01	0.23	-0.03	**0.55**
Time between identifying the artery and placing the vessel loop	-0.21	**0.64**	0.18	-0.25
Environment adjustment (workspace)	0.07	**0.53**	0.15	0.01
Environment adjustment (self)	0.12	-0.20	0.39	0.11
Environment adjustment (patient)	0.15	0.43	-0.17	0.26
Repositioning retractors	0.30	-0.22	0.08	0.20
Placing retractors or holding the incision open	**0.52**	-0.06	-0.10	-0.21
Laying out instruments ahead of time	-0.15	-0.04	0.11	-0.08
Extending the incision	0.42	0.12	0.37	0.09
Proportion of the time using instruments in two hands during the incision phase	-0.03	0.09	**0.67**	0.06
Proportion of the time using instruments in two hands during the muscle phase	-0.09	0.25	**0.59**	0.31
Proportion of the time using instruments in two hands during the identification phase	-0.25	0.11	0.45	0.32
Proportion of the time using instruments in two hands during the control phase	-0.22	0.23	0.29	0.13
Total instrument changes	0.24	**0.81**	0.23	-0.08
Instrument changes during the incision phase	0.17	0.04	**0.78**	-0.14
Instrument changes during the muscle phase	-0.05	**0.68**	0.03	-0.11
Instrument changes during the identification phase	0.48	0.45	0.02	0.06
Instrument changes during the control phase	-0.14	**0.55**	0.11	-0.03
Shifts between blunt and sharp dissection	0.17	-0.24	**0.63**	0.28
Proportion of the time sharp dissection used during the incision phase	-0.06	0.06	-0.31	0.28
Proportion of the time sharp dissection used during the muscle phase	0.06	0.02	0.13	**0.83**
Proportion of the time sharp dissection used during the identification phase	-0.08	-0.09	0.17	**0.58**
Proportion of the time sharp dissection used during the control phase	-0.12	0.14	0.42	-0.05
Proportion of the time blunt dissection used during the muscle phase	-0.06	-0.02	-0.13	-0.83
Proportion of the time blunt dissection used during the identification phase	0.11	0.12	-0.12	-0.52
Time spent searching for instruments	0.30	**0.66**	0.28	0.08
Idle time	0.50	0.45	0.51	0.19
Exploration	0.56	-0.14	0.52	0.01
Evaluating structures (by feel)	0.47	-0.05	0.37	-0.11
Checking by feel	0.35	0.46	-0.13	0.08
Evaluator assistance	0.57	0.39	-0.14	-0.07
Backtracking	0.59	0.08	0.40	0.17
Realizing a mistake	**0.70**	0.09	-0.15	0.15

3 factor model.

Rotated factor loadings for the model retaining three factors are listed in Table E.8 below. Factor 1 included *Weighing options, Workarounds, Evaluating structures (behavior), Heuristics,* and *Realizing a mistake.* Factor 2 contained *Time between identifying the artery and placing the vessel loop, Environment adjustment (workspace), Total instrument changes, Instrument changes during the muscle phase, Instrument changes during the control phase,* and *Time spent searching for instruments.* Factor 3 consisted of *Proportion of the time using instruments in two hands during the incision and muscle phases, Instrument changes during the incision,* and *Shifts between blunt and sharp dissection.* Although these three factors were slightly more coherent than those in the four factor model, they did not capture the same range of behaviors as the model retaining five factors. The model retaining three factors was not analyzed further.

Table E.8

Rotated factor loadings for the three factor model.

Variable	Factor 1	Factor 2	Factor 3
Forms a plan	0.36	0.63	0.02
Weighing options	**0.51**	0.01	-0.19
Expressions of confidence or certainty	0.44	-0.19	0.14
Expressions of doubt or uncertainty	0.63	-0.07	0.35
Mentioning things they expect to happen	0.38	0.14	-0.03
Double checking behaviors	0.01	-0.07	0.04
Evaluating progress	0.53	0.37	-0.19
Risk mitigation	0.47	0.44	0.11
Accounting for individual anatomy	0.22	0.08	-0.13
Workarounds	**0.60**	0.21	0.12
Naming structures	0.28	0.02	0.21
Evaluating structures (location)	0.42	0.09	-0.05
Evaluating structures (behavior)	**0.53**	-0.14	0.04
Knowledge	0.18	0.09	-0.02
Heuristics	**0.62**	0.28	0.12
Balancing constraints	0.11	0.22	-0.15
Evaluator hint	0.35	0.03	0.13
Evaluator prompting	0.30	-0.08	0.12
Altering a vessel loop	0.30	-0.24	0.21
Miscellaneous oddities	0.19	0.18	0.17
Time between identifying the artery and placing the vessel loop	-0.25	**0.68**	-0.01
Environment adjustment (workspace)	0.09	**0.54**	0.10
Environment adjustment (self)	0.03	-0.16	0.43
Environment adjustment (patient)	0.29	0.37	-0.07
Repositioning retractors	0.29	-0.23	0.21
Placing retractors or holding the incision open	0.44	-0.07	-0.09
Laying out instruments ahead of time	-0.19	-0.01	0.05
Extending the incision	0.34	0.15	0.42
Proportion of the time using instruments in two hands during the incision phase	-0.17	0.17	**0.62**
Proportion of the time using instruments in two hands during the muscle phase	-0.12	0.30	**0.63**
Proportion of the time using instruments in two hands during the identification phase	-0.24	0.15	0.49
Proportion of the time using instruments in two hands during the control phase	-0.22	0.27	0.27
Total instrument changes	0.22	**0.83**	0.15
Instrument changes during the incision phase	-0.07	0.15	**0.68**
Instrument changes during the muscle phase	-0.02	**0.69**	-0.08
Instrument changes during the identification phase	0.50	0.42	0.07
Instrument changes during the control phase	-0.11	**0.57**	0.03
Shifts between blunt and sharp dissection	0.05	-0.18	**0.72**
Proportion of the time sharp dissection used during the incision phase	0.10	-0.01	-0.19
Proportion of the time sharp dissection used during the muscle phase	0.24	-0.04	0.44
Proportion of the time sharp dissection used during the identification phase	0.03	-0.12	0.37
Proportion of the time sharp dissection used during the control phase	-0.21	0.20	0.33
Proportion of the time blunt dissection used during the muscle phase	-0.24	0.04	-0.44
Proportion of the time blunt dissection used during the identification phase	0.01	0.15	-0.30
Time spent searching for instruments	0.29	**0.67**	0.27
Idle time	0.43	0.47	0.57
Exploration	0.39	-0.09	0.57
Evaluating structures (by feel)	0.32	-0.01	0.36
Checking by feel	0.42	0.41	-0.08
Evaluator assistance	0.58	0.35	-0.11
Backtracking	0.51	0.10	0.50
Realizing a mistake	**0.74**	0.03	0.00

2 factor model.

Rotated factor loadings for the model retaining two factors are found in Table E.9 below. Factor 1 contained *Forms a plan, Evaluating progress, Risk mitigation, Workarounds, Heuristics, Total instrument changes, Instrument changes during the identification phase, Time spent searching for instruments, Checking by feel, Evaluator assistance,* and *Realizing a mistake.* Factor 2 included *Proportion of the time using instruments in two hands during the incision, Instrument changes during the incision, Shifts between blunt and sharp dissection,* and *Exploration.* Factor 1 appeared to lack coherence as it contained variables related to both planning/metacognitive processes and technical processes such as instrument use. This model was therefore rejected.

Table E.9

Rotated factor loadings for the two factor model.

Variable	Factor 1	Factor 2
Forms a plan	**0.69**	0.00
Weighing options	0.39	-0.03
Expressions of confidence or certainty	0.18	0.29
Expressions of doubt or uncertainty	0.38	0.52
Mentioning things they expect to happen	0.37	0.05
Double checking behaviors	-0.04	0.05
Evaluating progress	**0.65**	-0.10
Risk mitigation	**0.63**	0.15
Accounting for individual anatomy	0.22	-0.08
Workarounds	**0.57**	0.24
Naming structures	0.20	0.27
Evaluating structures (location)	0.37	0.05
Evaluating structures (behavior)	0.29	0.22
Knowledge	0.19	0.02
Heuristics	**0.63**	0.23
Balancing constraints	0.24	-0.15
Evaluator hint	0.26	0.21
Evaluator prompting	0.15	0.21
Altering a vessel loop	0.03	0.33
Miscellaneous oddities	0.25	0.17
Time between identifying the artery and placing the vessel loop	0.30	-0.21
Environment adjustment (workspace)	0.42	0.02
Environment adjustment (self)	-0.12	0.44
Environment adjustment (patient)	0.47	-0.06
Repositioning retractors	0.03	0.32
Placing retractors or holding the incision open	0.27	0.05
Laying out instruments ahead of time	-0.15	0.00
Extending the incision	0.32	0.46
Proportion of the time using instruments in two hands during the incision phase	-0.05	**0.51**
Proportion of the time using instruments in two hands during the muscle phase	0.07	0.50
Proportion of the time using instruments in two hands during the identification phase	-0.11	0.37
Proportion of the time using instruments in two hands during the control phase	0.01	0.14
Total instrument changes	**0.72**	0.04
Instrument changes during the incision phase	0.00	**0.60**
Instrument changes during the muscle phase	0.47	-0.21
Instrument changes during the identification phase	**0.64**	0.12
Instrument changes during the control phase	0.31	-0.12
Shifts between blunt and sharp dissection	-0.14	**0.73**
Proportion of the time sharp dissection used during the incision phase	0.08	-0.15
Proportion of the time sharp dissection used during the muscle phase	0.12	0.49
Proportion of the time sharp dissection used during the identification phase	-0.09	0.38
Proportion of the time sharp dissection used during the control phase	-0.04	0.22
Proportion of the time blunt dissection used during the muscle phase	-0.12	-0.49
Proportion of the time blunt dissection used during the identification phase	0.13	-0.31
Time spent searching for instruments	**0.65**	0.21
Idle time	0.60	0.57
Exploration	0.18	**0.66**
Evaluating structures (by feel)	0.20	0.43
Checking by feel	**0.59**	-0.03
Evaluator assistance	**0.67**	-0.01
Backtracking	0.39	0.60
Realizing a mistake	**0.55**	0.20

1 factor model.

Factor loadings for the model containing one factor are listed in Table E.10 below. The one factor model contained *Forms a plan, Expressions of doubt or uncertainty, Evaluating progress, Risk mitigation, Workarounds, Heuristics, Extending the incision, Total instrument changes, Instrument changes during the identification phase, Time spent searching for instruments, Idle time, Checking by feel, Evaluator assistance, Backtracking,* and *Realizing a mistake.* This factor seemed to contain variables representing metacognitive processes, processes related to adaptation, and technical skills. This lack of consistency caused us to reject the model containing only one factor.

Table E.10

Factor loadings for the one factor model.

Variable	Factor 1
Forms a plan	**0.60**
Weighing options	0.32
Expressions of confidence or certainty	0.30
Expressions of doubt or uncertainty	**0.59**
Mentioning things they expect to happen	0.34
Double checking behaviors	-0.01
Evaluating progress	**0.51**
Risk mitigation	**0.62**
Accounting for individual anatomy	0.15
Workarounds	**0.61**
Naming structures	0.31
Evaluating structures (location)	0.34
Evaluating structures (behavior)	0.35
Knowledge	0.18
Heuristics	**0.67**
Balancing constraints	0.13
Evaluator hint	0.33
Evaluator prompting	0.24
Altering a vessel loop	0.19
Miscellaneous oddities	0.30
Time between identifying the artery and placing the vessel loop	0.15
Environment adjustment (workspace)	0.38
Environment adjustment (self)	0.12
Environment adjustment (patient)	0.38
Repositioning retractors	0.19
Placing retractors or holding the incision open	0.26
Laying out instruments ahead of time	-0.13
Extending the incision	**0.50**
Proportion of the time using instruments in two hands during the incision phase	0.21
Proportion of the time using instruments in two hands during the muscle phase	0.31
Proportion of the time using instruments in two hands during the identification phase	0.09
Proportion of the time using instruments in two hands during the control phase	0.08
Total instrument changes	**0.64**
Instrument changes during the incision phase	0.30
Instrument changes during the muscle phase	0.30
Instrument changes during the identification phase	**0.62**
Instrument changes during the control phase	0.21
Shifts between blunt and sharp dissection	0.24
Proportion of the time sharp dissection used during the incision phase	-0.01
Proportion of the time sharp dissection used during the muscle phase	0.35
Proportion of the time sharp dissection used during the identification phase	0.12
Proportion of the time sharp dissection used during the control phase	0.07
Proportion of the time blunt dissection used during the muscle phase	-0.35
Proportion of the time blunt dissection used during the identification phase	-0.05
Time spent searching for instruments	**0.67**
Idle time	**0.80**
Exploration	0.48
Evaluating structures (by feel)	0.39
Checking by feel	**0.50**
Evaluator assistance	**0.58**
Backtracking	**0.64**
Realizing a mistake	**0.57**

APPENDIX F

ROTATED FACTOR LOADINGS FOR THE MODEL RETAINING FIVE FACTORS (ALL VARIABELS INCLUDED)

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Forms a plan	0.15	0.63	-0.04	0.27	0.35
Weighing options	0.29	0.01	-0.25	0.12	0.47
Expressions of confidence or certainty	0.42	-0.19	0.06	0.06	0.23
Expressions of doubt or uncertainty	**0.76**	-0.07	0.25	-0.05	0.16
Mentioning things they expect to happen	0.21	0.12	0.00	-0.02	**0.53**
Double checking behaviors	0.01	-0.07	0.04	0.01	0.02
Evaluating progress	0.26	0.37	-0.22	0.12	0.55
Risk mitigation	0.31	0.44	0.02	0.24	0.35
Accounting for individual anatomy	0.05	0.11	-0.30	0.41	-0.04
Workarounds	0.49	0.24	-0.12	0.43	0.08
Naming structures	0.05	-0.03	0.27	0.10	**0.75**
Evaluating structures (location)	0.45	0.10	-0.09	-0.12	0.11
Evaluating structures (behavior)	0.38	-0.11	-0.19	0.45	0.10
Knowledge	-0.02	0.05	0.08	-0.04	**0.63**
Heuristics	0.58	0.32	-0.13	0.38	-0.05
Balancing constraints	-0.15	0.20	-0.08	0.06	**0.53**
Evaluator hint	0.40	0.04	0.03	0.06	-0.01
Evaluator prompting	0.31	-0.10	0.14	-0.11	0.29
Altering a vessel loop	0.29	-0.21	-0.01	0.41	-0.17
Miscellaneous oddities	0.02	0.20	-0.02	**0.57**	0.00
Time between identifying the artery and placing the vessel loop	-0.19	**0.67**	0.13	-0.21	-0.05
Environment adjustment (workspace)	0.12	**0.54**	0.08	0.05	-0.07
Environment adjustment (self)	0.11	-0.19	0.41	0.07	0.05
Environment adjustment (patient)	0.14	0.39	-0.19	0.30	0.08
Repositioning retractors	0.27	-0.23	0.10	0.18	0.11
Placing retractors or holding the incision open	0.48	-0.07	-0.11	-0.21	0.19
Laying out instruments ahead of time	-0.22	-0.04	0.16	-0.09	0.17
Extending the incision	0.52	0.15	0.29	0.09	-0.18
Proportion of the time using instruments in two hands during the incision phase	-0.02	0.13	**0.68**	0.03	-0.01
Proportion of the time using instruments in two hands during the muscle phase	-0.03	0.28	0.58	0.30	-0.14
Proportion of the time using instruments in two hands during the identification phase	-0.15	0.14	0.43	0.31	-0.28
Proportion of the time using instruments in two hands during the control phase	-0.05	0.28	0.20	0.15	-0.48
Total instrument changes	0.22	**0.82**	0.18	-0.03	0.16
Instrument changes during the incision phase	0.20	0.10	**0.77**	-0.17	0.00
Instrument changes during the muscle phase	-0.02	**0.69**	-0.03	-0.06	-0.04
Instrument changes during the identification phase	0.41	0.42	0.01	0.09	0.30
Instrument changes during the control phase	-0.17	**0.55**	0.10	0.00	0.10
Shifts between blunt and sharp dissection	0.15	-0.22	**0.68**	0.23	0.10
Proportion of the time sharp dissection used during the incision phase	-0.03	0.03	-0.33	0.30	-0.11
Proportion of the time sharp dissection used during the muscle phase	0.07	-0.02	0.16	**0.82**	-0.01
Proportion of the time sharp dissection used during the identification phase	-0.19	-0.14	0.28	0.55	0.30
Proportion of the time sharp dissection used during the control phase	-0.10	0.17	0.41	-0.06	-0.04
Proportion of the time blunt dissection used during the muscle phase	-0.07	0.02	-0.16	-0.82	0.01
Proportion of the time blunt dissection used during the identification phase	0.23	0.17	-0.24	-0.49	-0.32
Time spent searching for instruments	0.27	**0.66**	0.25	0.11	0.19
Idle time	0.54	0.47	0.45	0.20	0.02
Exploration	0.57	-0.11	0.51	-0.02	0.09
Evaluating structures (by feel)	0.52	-0.02	0.33	-0.13	-0.02
Checking by feel	0.45	0.46	-0.24	0.13	-0.20
Evaluator assistance	0.54	0.37	-0.18	-0.03	0.19
Backtracking	**0.68**	0.11	0.33	0.17	-0.13
Realizing a mistake	**0.72**	0.06	-0.20	0.16	0.08

APPENDIX G

ROTATED FACTOR LOADINGS FOR THE MODEL RETAINING FIVE FACTORS (ITERATION TWO AND THREE – VARIABLES THAT DID NOT LOAD ONTO A FACTOR REMOVED)

Iteration 2.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Expressions of doubt or uncertainty	-0.11	0.17	**0.77**	0.17	0.02
Backtracking	0.18	0.39	0.66	-0.03	0.05
Realizing a mistake	0.08	-0.24	**0.80**	0.00	0.15
Time between identifying the artery and placing the vessel loop	**0.72**	0.06	-0.27	-0.08	0.04
Environment adjustment (workspace)	0.52	-0.04	0.23	-0.04	-0.03
Total instrument changes	**0.89**	0.14	0.17	0.19	-0.04
Instrument changes during the muscle phase	**0.80**	-0.12	0.01	-0.03	-0.03
Instrument changes during the control phase	**0.58**	0.04	-0.26	0.19	0.12
Time spent searching for instruments	**0.76**	0.25	0.20	0.17	0.21
Proportion of the time using instruments in two hands during the incision phase	0.13	**0.83**	-0.08	0.01	0.07
Instrument changes during the incision phase	0.12	**0.82**	0.17	0.03	-0.31
Shifts between blunt and sharp dissection	-0.14	**0.70**	0.10	0.11	0.35
Miscellaneous oddities	0.17	-0.02	-0.05	0.02	**0.80**
Proportion of the time sharp dissection used during the muscle phase	-0.02	0.09	0.26	0.00	**0.74**
Mentioning things they expect to happen	0.06	0.12	0.20	**0.54**	-0.27
Naming structures	-0.02	0.25	0.06	**0.81**	0.14
Knowledge	-0.01	-0.03	0.04	**0.77**	0.00
Balancing constraints	0.16	-0.12	-0.09	**0.55**	0.07

Iteration 3.

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Expressions of doubt or uncertainty	-0.12	0.22	0.11	**0.75**	0.06
Realizing a mistake	0.08	-0.20	-0.06	**0.80**	0.20
Time between identifying the artery and placing the vessel loop	**0.70**	0.05	-0.06	-0.31	0.05
Total instrument changes	**0.91**	0.13	0.13	0.20	-0.05
Instrument changes during the muscle phase	**0.82**	-0.14	-0.08	0.05	-0.04
Instrument changes during the control phase	**0.59**	0.00	0.21	-0.28	0.10
Time spent searching for instruments	**0.80**	0.24	0.11	0.23	0.19
Proportion of the time using instruments in two hands during the incision phase	0.13	**0.84**	0.00	-0.09	0.07
Instrument changes during the incision phase	0.13	**0.82**	-0.01	0.12	-0.30
Shifts between blunt and sharp dissection	-0.12	**0.71**	0.11	0.08	0.35
Miscellaneous oddities	0.18	-0.01	0.05	-0.03	**0.79**
Proportion of the time sharp dissection used during the muscle phase	-0.01	0.09	0.02	0.24	**0.74**
Mentioning things they expect to happen	0.14	0.15	0.41	0.40	-0.34
Naming structures	0.02	0.26	**0.80**	0.14	0.08
Knowledge	0.01	-0.02	**0.77**	0.11	-0.04
Balancing constraints	0.10	-0.11	**0.64**	-0.17	0.08

APPENDIX H

RESULTS FOR MARYLAND IPS ANALYSES

The section of the main text to which these analyses correspond is indicated for each analysis.

Section 4.2.1.1. Variance accounted for by WSU variables alone

I ran a regression model predicting Maryland IPS scores using scores on the five WSU factors. The WSU variables accounted for significant variance in IPS scores (R^2 = 0.38, F(5,63) = 7.83, p < 0.01). Scores on the strategy factor (b = -0.21, t(63) = -2.05, p = 0.04) and monitoring factor (b = -0.56, t(63) = -5.57, p < 0.01) both significantly predicted global outcome scores. The results of this analysis differ from the findings when using the WSU variables to predict Maryland global scores, because scores on the instrument change factor were no longer significant predictors here.

Section 4.2.1.2. Variance accounted for by WSU variables, controlling for Maryland variables

As with the analysis for the Maryland global score, I examined whether the WSU variables predicted variance in Maryland IPS scores beyond that accounted for by Maryland predictors. I ran a two-step regression predicting IPS scores with the Maryland variables in the first step and the WSU variables in the second step. Model one indicated that the Maryland variables significantly predicted IPS scores (F (7, 60) = 92.88, p < 0.01). Model two with the WSU variables included also predicted IPS scores (F (12, 55) = 51.51, p < 0.01). The model with only Maryland variables included accounted for

91.6% of the variance in IPS scores, while the model with the WSU variables included accounted for 91.8% of the variance in IPS scores. This change in R-squared of 0.3% was not statistically significant (*p* = 0.86).

Within model one, all of the Maryland variables predicted IPS scores. When the WSU factor scores were added to the model, all of the Maryland variables remained significant predictors. None of the WSU factor scores significantly predicted IPS scores. See table H.1 below for the full list of coefficients. Given the nature of the IPS score (i.e., it was calculated explicitly using the Maryland predictors as inputs), this result is not particularly surprising and I do not view it as discounting the results described in the main body of the paper.

Table H.1

Coefficients for models predicting IPS scores using Maryland and WSU variables.

Model	Variable	B	*t*	*p*
1	Q1 (suspected injury)	0.15	3.81	< 0.01
	Q3 (additional studies)	0.16	3.89	< 0.01
	Q7 (landmarks and incision)	0.18	3.94	< 0.01
	Q8S1 (steps of the procedure)	0.22	4.29	< 0.01
	Q8S2 (technique)	0.55	11.79	< 0.01
	Q9 (expert operative field maneuvers)	0.20	4.24	< 0.01
	Q12 (pitfalls)	0.10	2.61	0.01
2	Q1 (suspected injury)	0.15	3.37	< 0.01
	Q3 (additional studies)	0.16	3.90	< 0.01
	Q7 (landmarks and incision)	0.18	3.60	< 0.01
	Q8S1 (steps of the procedure)	0.24	4.00	< 0.01
	Q8S2 (technique)	0.56	10.60	< 0.01
	Q9 (expert operative field maneuvers)	0.19	3.73	< 0.01
	Q12 (pitfalls)	0.09	2.14	0.04
	Instrument change factor	0.03	0.67	0.51
	Strategy factor	0.02	0.53	0.60
	Declarative knowledge factor	-0.02	-0.36	0.72
	Problem identification factor	-0.01	-0.13	0.89
	Oddities factor	0.30	0.77	0.45

Section 4.2.2. Examining mediation between WSU and Maryland variables

I examined whether the Maryland variables mediated the relationship between WSU and IPS scores, just as I did with the Maryland global scores. Three of the mediation criteria described in the main body of the paper have been established in the previously described analyses. The relationship between WSU and Maryland predictors was established in the main body of the paper. I can therefore conclude that as with the Maryland global score, the Maryland predictors mediate the relationship between the WSU predictors and Maryland IPS scores.

Section 5.2.1.1. Self-confidence is positively correlated with performance

I first correlated surgeons' pre-procedure confidence ratings in their ability to perform the procedure with the Maryland IPS score. This correlation was significant ($r(38) = 0.48$, $p < 0.01$), again indicating that the surgeons were able to predict their own performance.

Section 5.2.2. Confidence changed in response to information from the world

As with the Maryland global scores, I examined whether surgeons' confidence ratings corresponded to IPS scores. I followed the same procedure as described in the main body of the text. I used IPS scores to predict surgeons' post-procedure confidence in their ability to perform the procedure, controlling for their pre-procedure confidence. The overall model was significant ($R^2 = 0.49$, $F(2, 35) = 16.71$, $p < 0.01$). Pre-procedure confidence significantly predicted post-procedure confidence ($b = 0.44$, $t(35) = 3.16$, $p < 0.01$), as did IPS scores ($b = 0.36$, $t(35) = 2.58$, $p = 0.01$). This result aligns closely with the results from the main body of the paper.

Section 5.2.3.1. Awareness is consistent across levels of experience

I also tested for an interaction between procedural confidence ratings and the surgeons' years of experience to see if the relationship between confidence ratings and ratings of performance changed as a function of experience. I constructed a two-step regression model predicting IPS scores using pre-procedure confidence ratings and surgeons' years of experience in the first step, and the interaction term in the second step. The model predicting IPS scores using pre-procedure procedural confidence ratings and the surgeons' experience was significant (R^2 = 0.24, F(2, 35) = 5.42, p = 0.01). Pre-procedure procedural confidence predicted IPS scores (b = 0.41, t(35) = 2.66, p = 0.01), but career experience did not (b = 0.17, t(35) = 1.10, p = 0.28). The interaction term did not increase the variance accounted for (R^2 = 0.24, F(3,34) = 3.52, p = 0.03; F change (1,34) = 0.03, p = 0.86), and the interaction term was not a significant predictor (b = -0.04, t(34) = -0.17, p = 0.86). These results mirror those in the main body of the paper.

Section 5.2.3.2. Awareness is consistent across levels of performance

As with experience, I next explored whether better performers were better attuned to their own performance based on interactions between the surgeons' self-reported confidence in their ability to perform the procedure and their performance tier (novice, journeyman, or expert). I constructed a model for the Maryland IPS scores using the surgeons' self-rated confidence in their ability to perform the exposure in the first step, then the interaction term between confidence and performance tier in the second step (performance tier was omitted from these models as it was inherently correlated with outcome scores). The model predicting Maryland IPS scores using pre-procedure procedural confidence ratings was significant (R^2 = 0.23, F(1, 38) = 11.61, p < 0.01).

Pre-procedure procedural confidence predicted Maryland IPS scores (b = 0.48, t(38) = 3.41, p < 0.01). The interaction term did not increase the variance accounted for (R^2 = 0.24, F(2,37) = 5.68, p = 0.01; F change (1,37) = 0.04, p = 0.85), and the interaction term was not a significant predictor (b = 0.11, t(37) = 0.19, p = 0.85). These findings again mirror those described in the body of the paper.

Section 6.2.1. Accounting for global scores using experience

6.2.2.1 Full data set. As with my main analysis, I first correlated the surgeons' years of experience with the IPS score. Years of experience was positively associated with Maryland IPS scores (r(84) = 0.35, p < 0.01), similar to results for the Maryland global scores.

I ran a series of regression models to determine how well experience accounted for variance in the Maryland IPS scores when controlling for training status. I first generated regression models predicting IPS scores that entered training status (pre-ASSET, post-ASSET, or expert) in step 1 and surgeons' years of experience in step 2. When predicting IPS scores, training phase by itself accounted for a significant proportion of the variance (R^2 = 0.54, F(1, 84) = 100.38, p < 0.01). Training status was a significant predictor of IPS scores (b = 0.74, t(84) = 10.02, p < 0.01). Unlike with Maryland global scores, surgeons' years of experience did not account for a significant additional proportion of the variance when controlling for training status (R^2 = 0.55, F(2, 83) = 23.01, p < 0.01; F change (1, 83) = 1.48, p = 0.23). Years of experience did not predict surgeons' IPS scores when controlling for training status (b = -0.11, t(83) = -1.23, p = 0.23).

6.2.1.2 Post-ASSET residents vs. attending surgeons. I investigated the relationship between experience and the Maryland IPS scores using regression in a manner similar to that described above. Using only the post-ASSET and attending surgeon procedure data, I generated a regression model with years of experience predicting Maryland IPS scores. As with the Maryland global scores, I noted a nonlinear trend in the residual plot. Unlike the analysis for the Maryland global scores, the overall quadratic model was significant ($R^2 = 0.23$, F (2, 44) = 6.51, $p < 0.01$).

Section 6.2.3.1. Prior training benefits performance

As in the main body of the paper, I used prior cadaver-based training (yes or no), total hours spent in the cadaver lab since medical school, and hours in the open skills lab since medical school as predictors to determine whether more specific experience was better able to predict performance than raw career experience. I generated regression models predicting the Maryland IPS measure using training status in the first step and training or experience prior to ASSET training in the second step.

When predicting Maryland IPS scores, training phase by itself accounted for a significant proportion of the variance ($R^2 = 0.53$, $F(1, 85) = 97.00$, $p < 0.01$). Training status was a significant predictor of Maryland IPS scores ($b = 0.73$, $t(85) = 9.85$, $p < 0.01$). When whether the surgeons had taken cadaver based courses prior to ASSET, the number of hours the surgeon had spent in the cadaver lab since medical school, and the number of hours the surgeon had spent in the open skills lab since medical school were added to the model, the resulting model was significant and accounted for significant additional variance in Maryland IPS scores compared to model 1 ($R^2 = 0.60$, $F(4, 82) = 30.38$, $p < 0.01$; F change (3, 82) = 4.35, $p = 0.01$). As with Maryland global scores, whether the surgeon had taken other cadaver-based courses before ASSET training significantly predicted Maryland IPS scores controlling for training status (*b* = 0.23, *t*(82) = 3.23, *p* < 0.01). The number of hours spent in the cadaver and open skills lab since medical school were not significant predictors (see Table H.2).

Table H.2
Predicting IPS scores using prior training/experience.

Model	Variable	B	*t*	*p*
1	Training phase	0.73	9.85	< 0.01
2	Training phase	0.67	8.49	< 0.01
	Prior cadaver-based courses	0.23	3.23	< 0.01
	Hours in the cadaver lab	0.37	1.53	0.13
	Hours in the open skills lab	-0.41	-1.74	0.09

APPENDIX I

RESULTS FOR WSU OBJECTIVE RANK SCORES

The section of the main text to which these analyses correspond is indicated for each analysis.

Section 4.2.1.1. Variance accounted for by WSU variables alone

As with the Maryland global scores, the assumption of normality in the residuals was violated for the WSU objective rank scores. We did not transform the data for the same reasons described in the main body of the text. Following the procedure for the other two outcome measures, I ran a regression using the five WSU factors to predict WSU objective rank scores. The WSU variables captured significant variance in WSU objective rank scores ($R^2 = 0.33$, $F(5, 63) = 6.09$, $p < 0.01$). Scores on the strategy factor ($b = -0.30$, $t(63) = -2.84$, $p = 0.01$) and monitoring factor ($b = -0.38$, $t(63) = -3.56$, $p < 0.01$) both significantly predicted global outcome scores. These findings replicate those for the IPS scores, but these analyses once again fail to replicate the finding from the Maryland global score analysis that scores on the instrument change factor predicted the outcome measure.

Section 4.2.1.2. Variance accounted for by WSU variables, controlling for Maryland variables

Similarly to the analyses for Maryland global and IPS scores, I explored whether the WSU factor scores accounted for variance in the WSU objective rank scores beyond that accounted for by the Maryland predictors. I predicted WSU objective rank scores using a two stage model with the Maryland predictors in the first step and WSU factor scores in the second step. Model one indicated that the Maryland variables significantly predicted WSU objective rank scores ($F(7,60) = 18.86$, $p < 0.01$). Model two with the WSU variables included also predicted these scores ($F(12,55) = 12.41$, $p < 0.01$). The model with only Maryland variables included accounted for 68.7% of the variance in WSU objective rank scores, while the model with the WSU variables included accounted for 73% of the variance in WSU objective rank scores. This change in R-squared of 4.3% was not statistically significant ($p = 0.86$).

Within model one, Q3 (additional studies), Q8S1 (procedure steps), and Q9 (expert operative field maneuvers) all predicted WSU objective rank scores (Q3 $B = 0.29$, $p < 0.01$; Q8S1 $B = 0.69$, $p < 0.01$; Q9 $B = 0.23$, $p = 0.02$). Within model two, the same Maryland variables remained significant predictors (Q3 $B = 0.27$, $p < 0.01$; Q8S1 $B = 0.59$, $p < 0.01$; Q9 $B = 0.28$, $p < 0.01$). None of the WSU variables predicted performance on the objective rank scores.

Although the additional variance accounted for by the WSU factors was similar between the Maryland global scores and WSU objective rank scores, this analysis failed to replicate the findings from the main body of the text. Such a finding may imply that the Maryland global scores rely on different criteria than the other measures. Because the global score relies on evaluator judgment and the other two measures are calculated somewhat more objectively, it is possible that the evaluators account for features of performance not included in the other measures.

Section 4.2.2. Examining mediation between WSU and Maryland variables

I examined whether the Maryland variables mediated the relationship between WSU factor scores and WSU objective rank scores, just as I did with the other outcome variables. Three of the criteria for mediation described in the main body of the paper have been established in the previously described analyses. The relationship between WSU and Maryland predictors was established in the main body of the paper. I can therefore conclude that as with the Maryland global score, the Maryland predictors mediate the relationship between the WSU predictors and WSU objective rank scores.

Section 5.2.1.1. Self-confidence is positively correlated with performance

I first correlated surgeons' pre-procedure confidence ratings in their ability to perform the procedure with the WSU objective rank score. This correlation was significant ($r(38) = 0.50, p < 0.01$), again indicating that the surgeons were able to predict their own performance.

Section 5.2.2 Confidence changed in response to information from the world

As with the other two outcome measures, I examined whether surgeons' confidence ratings corresponded to WSU objective rank scores. I followed the same procedure as described in the main body of the text. I used WSU objective rank scores to predict surgeons' post-procedure confidence in their ability to perform the procedure, controlling for their pre-procedure confidence. The overall model was significant ($R^2 = 0.48$, $F(2, 35) = 16.01$, $p < 0.01$). Pre-procedure confidence significantly predicted post-procedure confidence ($b = 0.43$, $t(35) = 2.98$, $p = 0.01$), as did the WSU objective rank score ($b = 0.35$, $t(35) = 2.41$, $p = 0.02$). This result matches the previous analyses with the other outcome measures.

Section 5.2.3.1. Awareness is consistent across levels of experience

I also tested for an interaction between procedural confidence ratings and the surgeons' years of experience to see if the relationship between confidence ratings and ratings of performance changed as a function of experience. I constructed a two-step regression model predicting WSU objective rank scores using pre-procedure confidence ratings and surgeons' years of experience in the first step, and the interaction term in the second step. Consistent with prior analyses, the first model predicting WSU objective rank scores using pre-procedure procedural confidence ratings and the surgeons' experience was significant (R^2 = 0.23, F(2, 35) = 5.30, p = 0.01). Pre-procedure procedural confidence predicted WSU objective rank scores (b = 0.50, t(35) = 3.24, p < 0.01), but career experience did not (b = -0.10, t(35) = -0.66, p = 0.52). The interaction term did not increase the variance accounted for (R^2 = 0.25, F(3,34) = 3.81, p = 0.02; F change (1,34) = 0.86, p = 0.36), and the interaction term was not a significant predictor (b = -0.19, t(34) = -0.93, p = 0.36).

Section 5.2.3.2. Awareness is consistent across levels of performance

As with experience, I next explored whether better performers were better attuned to their own performance based on interactions between the surgeons' self-reported confidence in their ability to perform the procedure and their performance tier (novice, journeyman, or expert). I constructed a model for the WSU objective rank scores using the surgeons' self-rated confidence in their ability to perform the exposure in the first step, then the interaction term between confidence and performance tier in the second step (performance tier was omitted from these models as it was inherently correlated with outcome scores). The model predicting WSU objective rank scores using pre-procedure procedural confidence ratings was significant (R^2 = 0.25, F(1, 38) = 12.75, p < 0.01). Pre-procedure procedural confidence predicted WSU objective rank scores (b = 0.50, t(38) = 3.57, p < 0.01). The interaction term did not increase the variance accounted for

(R^2 = 0.26, $F(2,37)$ = 6.47, $p < 0.01$; F change (1,37) = 0.39, p = 0.54), and the interaction term was not a significant predictor (b = 0.36, $t(37)$ = 0.63, p = 0.54). These findings are again consistent with prior analyses.

Section 6.2.1. Accounting for global scores using experience

6.2.1.1 Full data set. As with my main analysis, I first correlated the surgeons' years of experience with the WSU objective rank score. Unlike prior analyses, years of experience was not associated with Maryland IPS scores ($r(84)$ = 0.17, p = 0.12). I ran a series of regression models to determine how well experience accounted for variance in the WSU objective rank scores when controlling for training status. I first generated regression models predicting objective rank scores that entered training status (pre-ASSET, post-ASSET, or expert) in step 1 and surgeons' years of experience in step 2.

When predicting WSU objective rank scores, training phase by itself accounted for a significant proportion of the variance (R^2 = 0.30, $F(1, 84)$ = 35.76, $p < 0.01$). Training status was a significant predictor of WSU objective rank scores (b = 0.55, $t(84)$ = 5.98, $p < 0.01$). Surgeons' years of experience accounted for a significant additional proportion of the variance when controlling for training status (R^2 = 0.33, $F(2, 83)$ = 20.57, $p < 0.01$; F change (1, 83) = 4.07, p = 0.05). Both training status and years of experience predicted surgeons' WSU objective rank scores in the second model (b = 0.67, $t(83)$ = 6.14, $p < 0.01$ and b = -0.22, $t(83)$ = -2.02, p = 0.05, respectively), replicating the findings from the main body of the text.

6.2.1.2 Post-ASSET residents vs. attending surgeons. I investigated the relationship between experience and the Maryland IPS scores using regression in a manner similar to that described above. Using only the post-ASSET and attending surgeon procedure data, I generated a regression model with years of experience predicting WSU objective rank scores. As with the other outcome variables, the quadratic model best captured the variance in the WSU objective rank score ($R^2 = 0.14$, F (2, 44) = 3.53, $p = 0.04$).

*Section 6.2.3.1. Prior training did **not** benefit performance.*

As in the previous analyses for the other outcome measures, I used prior cadaver-based training (yes or no), total hours spent in the cadaver lab since medical school, and hours in the open skills lab since medical school as predictors to determine whether more specific experience was better able to predict performance than raw career experience. I generated regression models predicting the WSU objective rank measure using training status in the first step and training or experience prior to ASSET training in the second step.

When predicting WSU objective rank scores, training phase by itself accounted for a significant proportion of the variance ($R^2 = 0.28$, F(1, 85) = 32.23, $p < 0.01$). Training status was a significant predictor of WSU objective rank scores ($b = 0.53$, t(85) = 5.68, $p < 0.01$). When whether the surgeons had taken cadaver based courses prior to ASSET, the number of hours the surgeon had spent in the cadaver lab since medical school, and the number of hours the surgeon had spent in the open skills lab since medical school were added to the model, the resulting model was significant but failed to account for significant additional variance in WSU objective rank scores compared to model 1 ($R^2 = 0.32$, F(4, 82) = 9.75, $p < 0.01$; F change (3, 82) = 1.90, $p = 0.14$). Whether the surgeon had taken cadaver based courses prior to ASSET and the number of hours spent in the cadaver and open skills labs since medical school all failed to account for variance in WSU objective rank scores (see Table I.1). These results conflict with the other analyses, which indicated a benefit of cadaver-based courses prior to ASSET training.

Table I.1
Predicting objective rank scores using prior training/experience.

Model	Variable	B	*t*	*p*
1	Training phase	0.53	5.68	< 0.01
2	Training phase	0.51	4.98	< 0.01
	Prior cadaver-based courses	0.09	0.91	0.37
	Hours in the cadaver lab	0.44	1.40	0.17
	Hours in the open skills lab	-0.57	-1.87	0.07

APPENDIX J

REGRESSIONS PREDICTING GLOBAL SCORES, ALLOWING NONSIGNIFICANT PREDICTORS TO FALL OUT

Relevant sections in the main text are called out below.

Section 4.2.1.1. Variance accounted for by WSU variables alone

I entered all five WSU variables (scores on the instrument change factor, strategy factor, deliberate behavior factor, monitoring factor, and oddities factor) into regression models for the Maryland global outcome measure. Overall, WSU variables accounted for significant variance in global scores ($R^2 = 0.47$, $F(5, 63) = 11.16$, $p < 0.01$). Specific scores on the instrument change factor ($b = -0.22$, $t(63) = -2.33$, $p = 0.02$), strategy factor ($b = -0.26$, $t(63) = -2.80$, $p = 0.01$), and monitoring factor ($b = -0.54$, $t(63) = -5.74$, $p < 0.01$) all significantly predicted global outcome scores. Each of the WSU predictors was negatively related to Maryland global performance scores, indicating that better-performing surgeons displayed fewer *instrument change, strategy,* and *monitoring* behaviors. Scores on the deliberate behavior and oddities factors were not significant ($p = 0.20$ and $p = 0.68$, respectively).

I next ran a model predicting Maryland global scores using only the significant predictors from the first model (instrument change, strategy, and monitoring behaviors). The model was again significant ($F(3, 65) = 18.02$, $p < 0.01$). Specific scores on the instrument change factor ($b = -0.23$, $t(65) = -2.49$, $p = 0.02$), strategy factor ($b = -0.27$,

$t(65) = -2.92$, $p = 0.01$), and monitoring factor ($b = -0.54$, $t(65) = -5.89$, $p < 0.01$) all significantly predicted global outcome scores.

Section 4.2.1.2. Variance accounted for by WSU variables, controlling for Maryland variables

I selected the final set of Maryland variables aggregated from the evaluation form (described in Chapter 3): Q1 (suspected injury), Q3 (additional studies), Q7 (landmarks and incision), Q8S1 (steps of the procedure), Q8S2 (technique), Q9 (expert operative field maneuvers), and Q12 (pitfalls) and entered them into a model predicting Maryland global scores. This model indicated that the Maryland variables alone significantly predicted global outcome scores ($R^2 = 0.78$, $F(7, 60) = 29.58$, $p < 0.01$). Within the model, only Q8S1 (steps of the procedure; $B = 0.54$, $t(60) = 6.50$, $p < 0.01$) and Q8S2 (technique; $B = 0.52$, $t(60) = 6.80$, $p < 0.01$) significantly predicted global scores. These variables were retained in step one of a subsequent model, with the five WSU factors added in step two to examine additional variance accounted for.

Step one of this model was significant ($R^2 = 0.79$, $F(2, 66) = 123.27$, $p < 0.01$), as was step two ($R^2 = 0.82$, $F(7, 61) = 39.84$, $p < 0.01$). The change in R^2 between the two models was not significant (F change $(5,61) = 2.16$, $p = 0.07$). Within step two, only Q8S1 (steps of the procedure; $B = 0.45$, $t(61) = 6.19$, $p < 0.01$), Q8S2 (technique; $B = 0.47$, $t(61) = 6.64$, $p < 0.01$), and scores on the deliberate behavior factor ($B = -0.15$, $t(61) = -2.63$, $p = 0.01$) significantly predicted global scores. These surviving predictors were entered into a final model predicting Maryland global scores with the Maryland predictors in step one and the WSU predictor in step two (Table J.1 below).

Table J.1

Final model predicting Maryland global scores with Maryland and WSU variables.

Model	Variable	B	*t*	*p*
1	Q8S1 (steps of the procedure)	0.55	8.99 (87)	< 0.01
	Q8S2 (technique)	0.44	7.20 (87)	< 0.01
2	Q8S1 (steps of the procedure)	0.54	9.03 (86)	< 0.01
	Q8S2 (technique)	0.46	7.52 (86)	< 0.01
	Deliberate behavior factor	-0.09	-2.05 (86)	0.04

Section 6.2.3.1. Prior training benefits performance

I used prior cadaver-based training (yes or no), total hours spent in the cadaver lab since medical school, and total hours in the open skills lab since medical school as predictors to determine whether specific types of experience prior to ASSET training were able to predict performance. I generated regression models predicting the Maryland global outcome measure using training status (pre-ASSET, post-ASSET, or attending physician) in the first step and training or experience prior to ASSET training in the second step.

When predicting Maryland global scores, training phase by itself accounted for a significant proportion of the variance (R^2 = 0.43, F(1, 85) = 63.32, p < 0.01). Training status was a significant predictor of Maryland global scores (b = 0.65, t(85) = 7.96, p < 0.01). I next added whether the surgeons had taken cadaver based courses prior to ASSET and the number of hours the surgeon had spent in the cadaver or open skills labs since medical school to the model. The resulting model was significant but did not predict significant additional variance in Maryland global scores compared to model 1 (R^2 = 0.47, F(4, 82) = 18.20, p < 0.01; F change (3, 82) = 2.23, p = 0.09). Despite the lack of significant additional variance accounted for by the additional variables overall, whether the surgeon had taken other cadaver-based courses before ASSET training did significantly predict Maryland global scores controlling for training status (b = 0.19, t(82)

= 2.31, *p* = 0.02). The number of hours spent in the cadaver and open skills lab since medical school were not significant predictors (Table J.2).

Table J.2
Predicting Maryland global scores using training status and other types of experience.

Model	Variable	B	*t*	*p*
1	Training phase	0.65	7.96 (88)	< 0.01
2	Training phase	0.64	7.13 (85)	< 0.01
	Prior cadaver-based courses	0.19	2.31 (85)	0.02
	Hours in the cadaver lab	-0.01	-0.02 (85)	0.99
	Hours in the open skills lab	-0.09	-0.32 (85)	0.75

I next generated a model using only the significant predictors from above, namely, training phase (step 1) and prior cadaver-based courses (step 2). Both predictors in the resulting model remained significant when the prior nonsignificant variables were excluded (Table J.3).

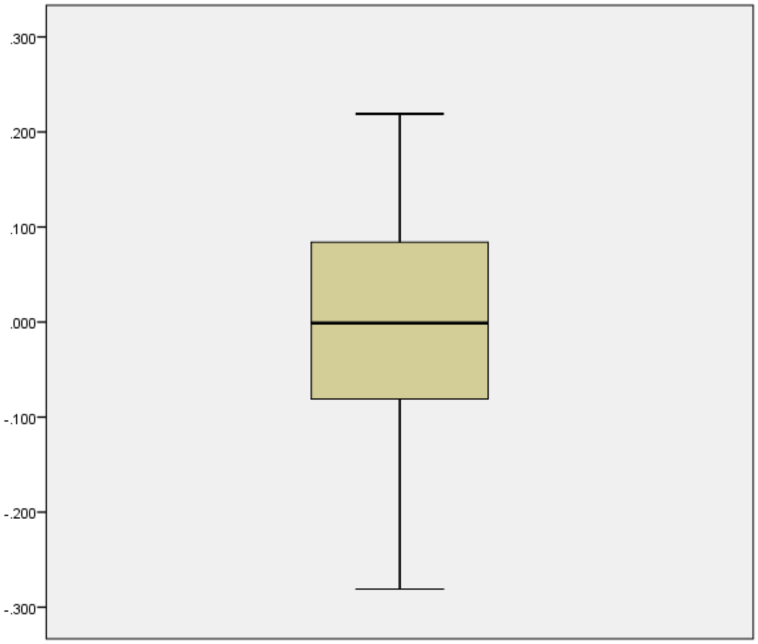
Table J.3
Final model predicting Maryland global scores using only significant prior training predictors.

Model	Variable	B	*t*	*p*
1	Training phase	0.65	7.96 (88)	< 0.01
2	Training phase	0.62	7.57 (87)	< 0.01
	Prior cadaver-based courses	0.19	2.37 (87)	0.02

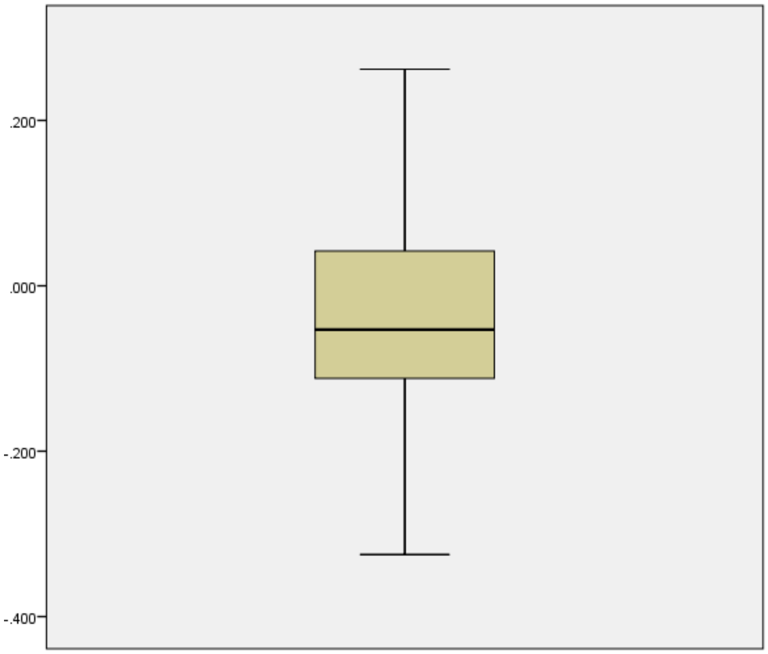
APPENDIX K

BOX AND WHISKER PLOTS FOR REMAINING LAGS OF THE TIME SERIES ANALYSIS EXAMINING THE INFLUENCE OF ITEM ORDER ON EVALUATIONS

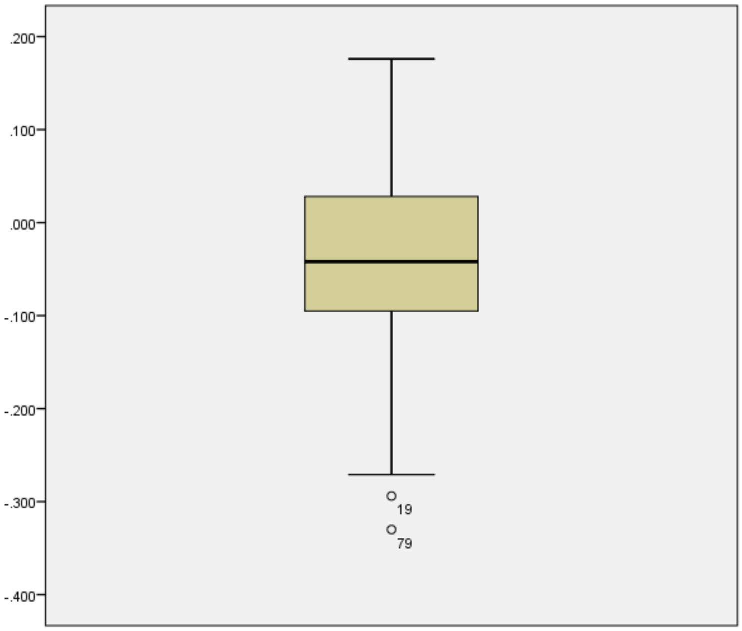
Lag 1.



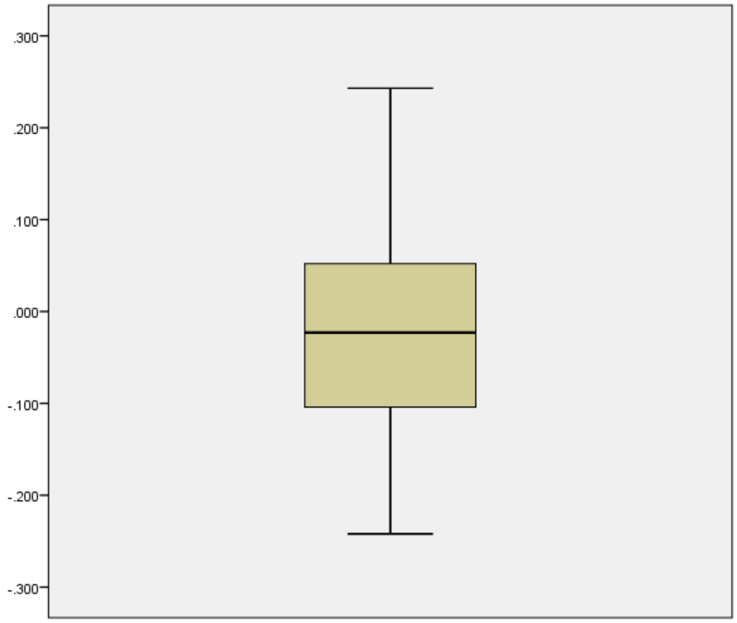
Lag 3.



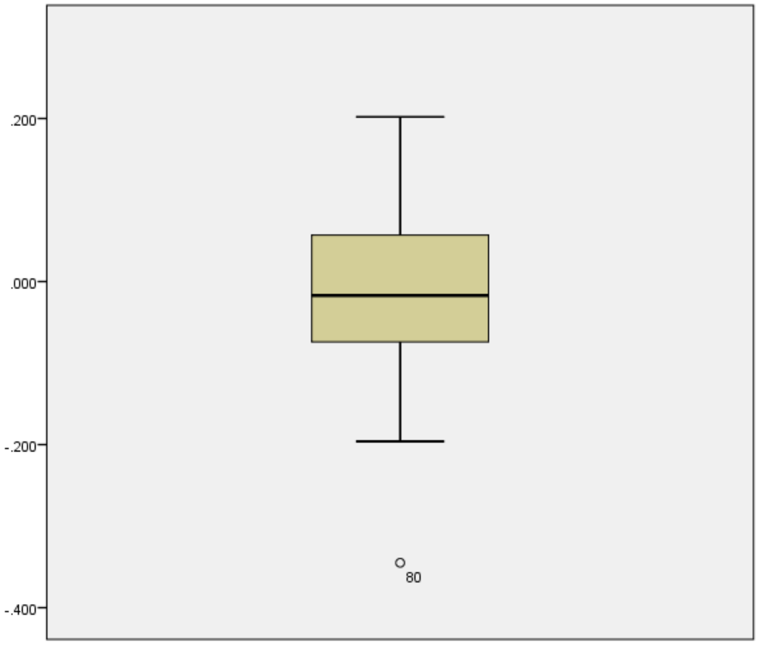
Lag 4.



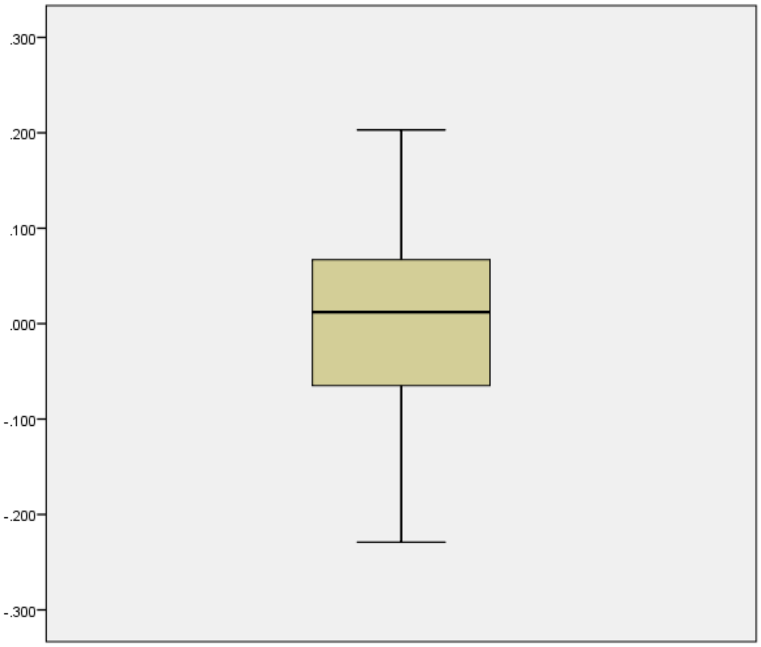
Lag 5.



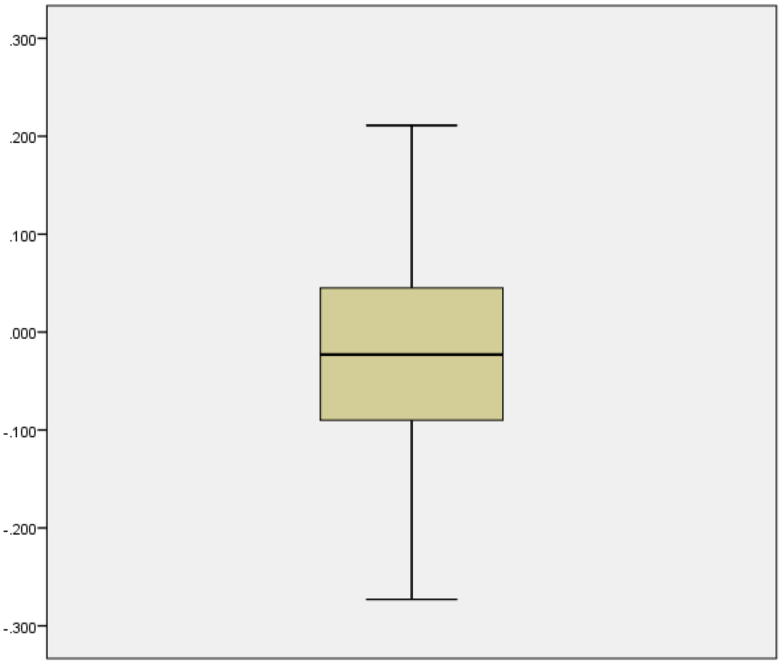
Lag 6.



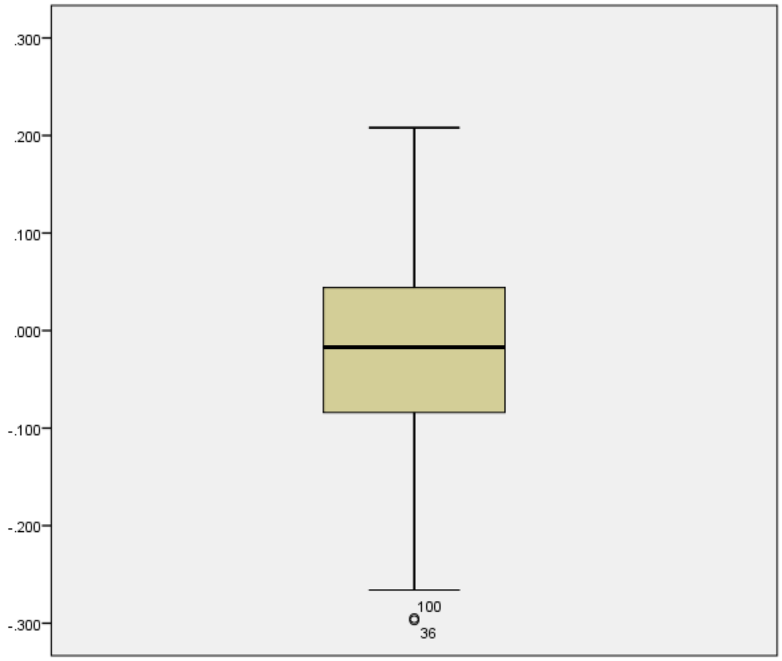
Lag 7.



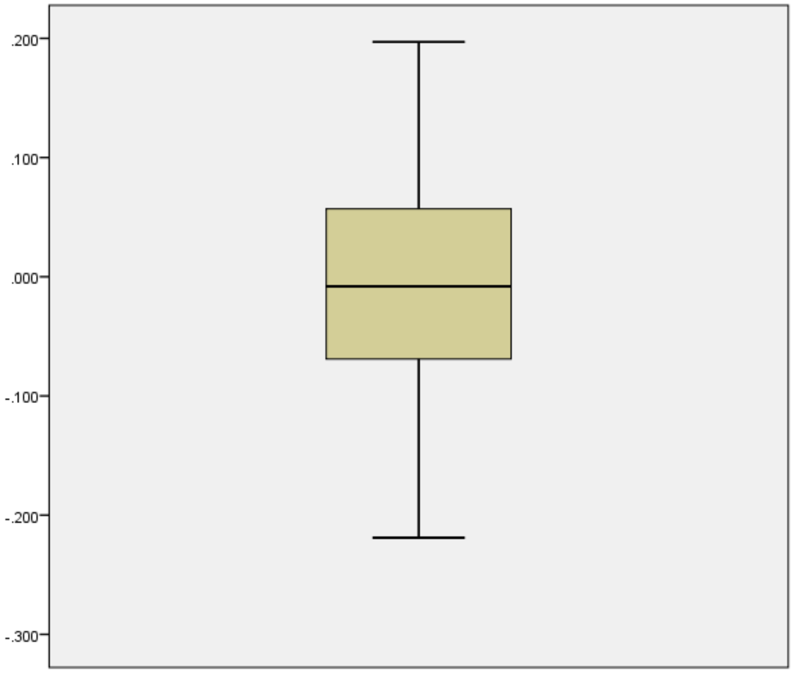
Lag 8.



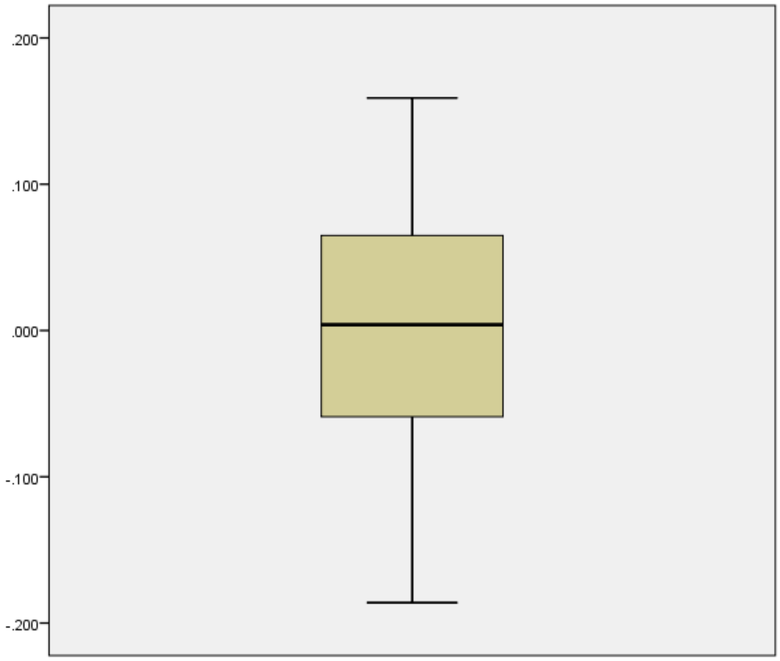
Lag 9.



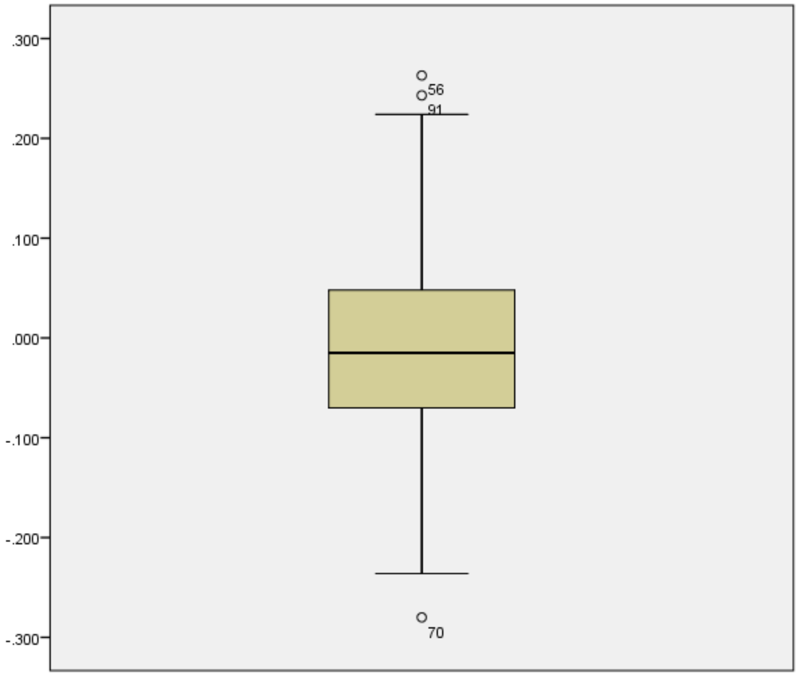
Lag 10.



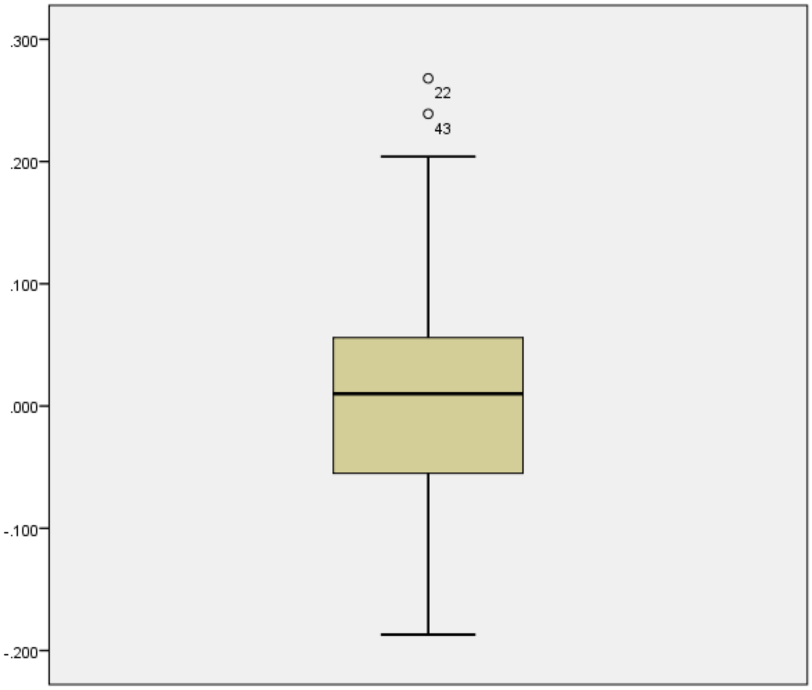
Lag 11.



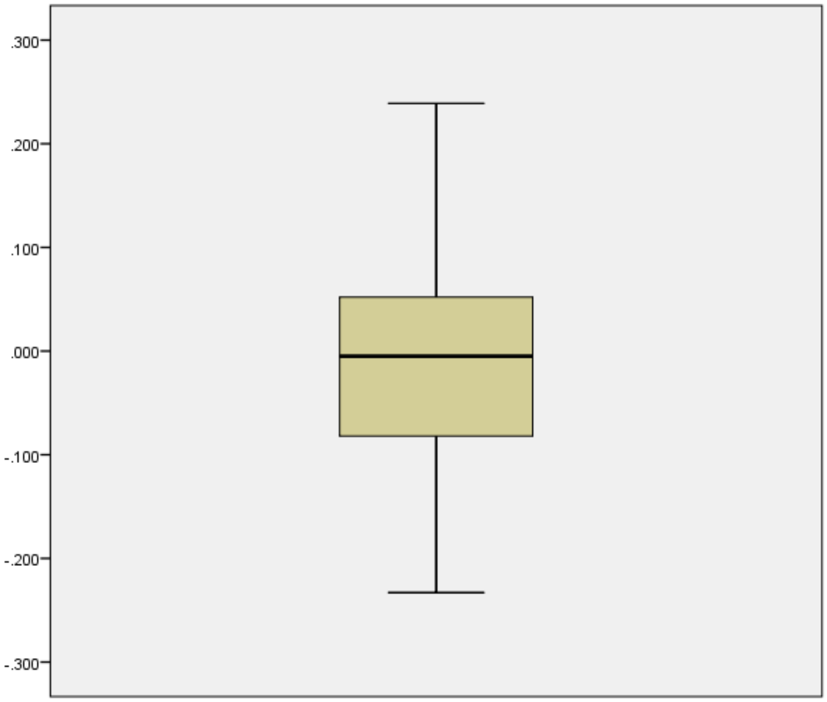
Lag 12.



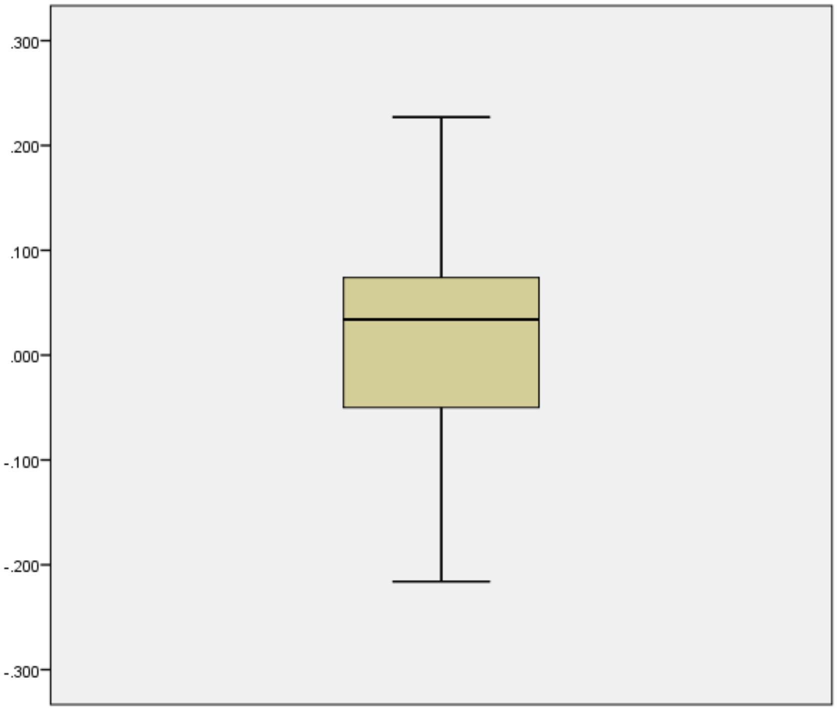
Lag 13.



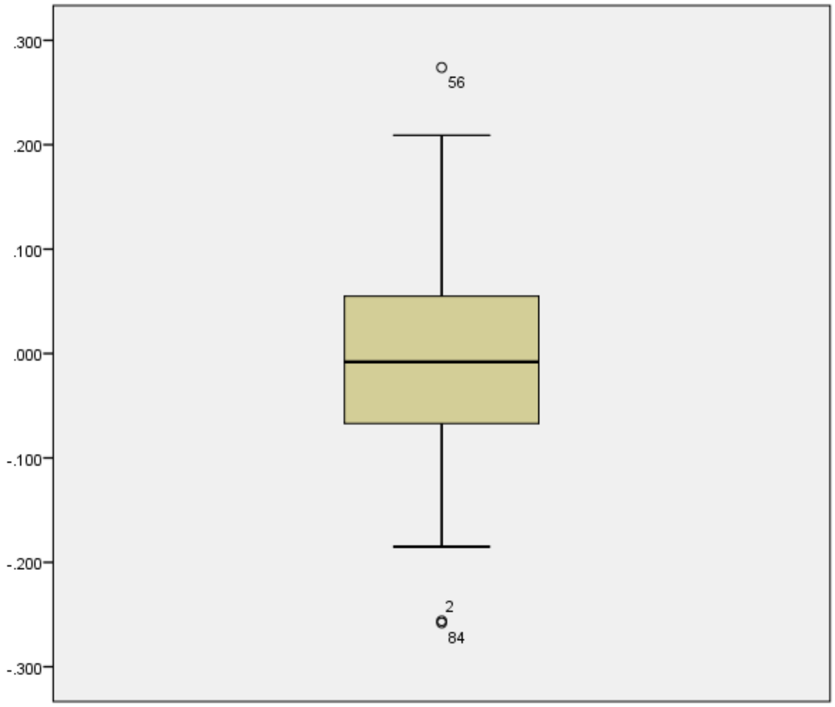
Lag 14.



Lag 15.



Lag 16.



References

Abernethy, B., Poolton, J., Masters, R., & Patil, N. (2008). Implications of an expertise model for surgical skills training. *ANZ J Surg., 78,* 1092-1095.

Ahmed, K., Miskovic, D., Darzi, A., Athanasiou, T., & Hanna, G. (2011). Observational tools for assessment of procedural skills: A systematic review. *The American Journal of Surgery, 202,* 469-480.

Alderson, D. (2010). Developing expertise in surgery. *Medical Teacher, 32,* 830-836.

Anderson, J. (1992). Automaticity and the ACT* theory. *American Journal of Psychology, 105,* 165-180.

Anderson, J., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction, 12,* 439-462.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics, 34,* 555-596.

Balk, Y., Adriaanse, M., de Ridder, D., & Evers, C. (2013). Coping under pressure: Employing emotion regulation strategies to enhance performance under pressure. *Journal of Sport & Exercise Physiology, 35,* 408-418.

Balzer, W., & Sulsky, L. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77,* 975-985.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Baumeister, R. (1984). Choking under pressure: Self-consciousnes and paradoxical effects of incentives on skillful performance. *Journal of Personality & Social Psychology, 46,* 610-620.

Bech, B., Lonn, L., Schroeder, T., Rader, S., & Ringsted, C. (2010). Capturing the essence of developing endovascular expertise for the construction of a global assessment instrument. *European Journal of Vascular & Endovascular Surgery, 40,* 292-302.

Beilock, S., Bertenthal, B., McCoy, A., & Carr, T. (2004). Haste does not always make waste: Expertise, direction of attention, and speed versus accuracy in performing sensorimotor skills. *Psychonomic Bulletin & Review, 11,* 373-379.

Bliese, P. (2002). Multilevel random coefficient modeling in organizational research: Examples using SAS and S-Plus. In F. Drazgow & N. Schmitt (Eds.), Measuring and analyzing behavior in organizations: Advances in measurement and data analysis (pp. 401-445). San Francisco, CA: Jossey-Bass, Inc.

Borman, W. (1987). Personal constructs, performance schemata, and "folk theories" of subordinate effectiveness: Explorations in an Army officer sample. *Organizational Behavior and Human Decision Processes, 40,* 307-322.

Bradley, J.H., Paul, R., & Seeman, E. (2006). Analysing the structure of expert knowledge. *Information and Management, 43,* 77-91.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152.

Cohen, J., Cohen, P., West, S., & Aiken, L. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Clancey, W. Task models vs work practice simulations. *Unpublished document.*

DeGroot, A. (1965). Thought and choice in chess. Amsterdam Academic Archive.

De Swert, N. (2012). Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. *Retrieved from: http://www.polcomm.org/wp-content/uploads/ICR01022012.pdf*

Dennis, I. (2007). Halo effects in grading student projects. *Journal of Applied Psychology, 92,* 1169-1176.

Djulbegovich, B., Beckstead, J., Elqayam, S., Reljic, T., Hozo, I., Kumar, A., Canon-Bowers, J., et al., (2014). Evaluation of physicians' cognitive styles. *Medical Decision Making, 34,* 627-637.

Dreyfus, H.L., & Dreyfus, S.E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York, NY: The Free Press.

DuBois, D. & Shalin, V.L. (1995). Adapting cognitive methods to real-world objectives: An application to job knowledge testing. In P.D. Nichols, S.F. Chipman, R.L. Brennan (Eds.) Cognitively diagnostic assessment. Hillsdale, NJ: Lawrence Erlbaum Associates.

Duncker, K., (1926). A qualitative (experimental and theoretical) study of productive thinking (solving of comprehensible problems). *The Pedagogical Seminary and Journal of Genetic, 33,* 642-708.

Duncker, K. (1945). On problem-solving. *Psychological Monographs, 58(5),* 1-113.

Dunphy, B., & Williamson, S. (2004). In pursuit of expertise: Toward an educational model for expertise development. *Advances in Health Sciences Education, 9,* 107-127.

Ericsson, K. (2014). How to gain the benefits of the expert performance approach in domains where the correctness of decisions are not readily available: A reply to Weiss and Shanteau. *Applied Cognitive Psychology, 28,* 458-463.

Ericsson, K., Krampe, R., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363-406.

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.

Feeley, T. (2002). Comment on halo effects in rating and evaluation research. *Human Communication Research, 28,* 578-586.

Ferreira, M., Mata, A., Donkin, C., Sherman, S., & Ihmels, M. (2016). Analytic and heuristic processes in the detection and resolution of conflict. *Mem Cogn, 44,* 1050-1063.

Fisicaro, S., & Lance, C. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14,* 419-429.

Flach, J., & Warren, R. (1995). Active psychophysics: The relation between mind and what matters. In J. Flach, P. Hancock, J. Caird, and K. Vicente (eds.), Global Perspectives on the Ecology of Human Machine Systems. New Jersey: Lawrence Erlbaum Associates.

Flach, J., & Voorhorst, F., (2017). What matters? Flach & Voorhorst.

Gallagher, A., Richie, K., McClure, N., & McGuigan, J. (2001). Objective psychomotor skills assessment of experienced, junior, and novice laparoscopists with virtual reality. *World Journal of Surgery, 25,* 1478-1483.

Gelinas-Phaneuf, N., & Del Maestro, R. (2013). Surgical expertise in neurosurgery: Integrating theory into practice. *Neurosurgery,* 73, S30-S38.

Gibson, J., & Gibson, E. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review, 62,* 32-41.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3,* 20-29.

Greeno, J., & Moore, J. (1993). Situativity and symbols: Response to Vera and Simon. *Cognitive Science, 17,* 49-59.

Greeno, J., Riley, M., & Gelman, R. (1984). Conceptual competence and children's counting. *Cognitive Psychology, 16,* 94-143.

Hammond, K., & Summers, D. (1972). Cognitive control. *Psychological Review, 79,* 58-67.

Hayes, A., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1,* 77-89.

Hosking, S., Davey, C., & Kaiser, M. (2013). Visual cues for manual control of headway. *Frontiers in Behavioral Neuroscience, 7,* 1-14.

Huhn, J., Potts, C., & Rosenbaum, D. (2016). Cognitive framing in action. *Cognition, 151,* 42-51.

Hull, L., Arora, S., Aggarwal, R., Darzi, A., Vincent, C., & Sevdalis, N. (2012). The impact of nontechnical skills on technical performance in surgery: A systematic review. *J Am Coll Surg, 214,* 214-230.

Jackson, C., & Furnham, A. (2001). Appraisal ratings, halo, and selection: A study using sales staff. *European Journal of Psychological Assessment, 17,* 17-24.

Jelovsek, J., Kow, N., & Diwadkar, G. (2013). Tools for the direct observation and assessment of psychomotor skills in medical trainees: A systematic review. *Medical Education, 47,* 650-673.

Judd, C., Drake, R., Downing, J., & Krosnick, J. (1991). Some dynamic properties of attitude structures: Context-induced response facilitation and polarization. *Journal of Personality and Social Psychology, 60,* 193-202.

Juvina, I. (2011). Cognitive control: Componential and yet emergent. *Topics in Cognitive Science, 3,* 242-246.

Kaakinen, J., Hyona, J., & Viljanen, M. (2011). Influence of a psychological perspective on scene viewing and memory for scenes. *The Quarterly Journal of Experimental Psychology, 64,* 1372-1387.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18,* 513-549.

Klein, G.A. (1989). Recognition-primed decisions. *Advances in Man-Machine Systems Research, 5,* 47-92.

Kogan, J., Holmboe, E., & Hauer, K. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: A systematic review. *JAMA, 302,* 1316-1326.

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30,* 411-433.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-reliability. Retrieved from http://repository.upenn.edu/asc_papers/43.

Kulatunga-Moruzzi, C., Brooks, L., & Norman, G. (2004). Using comprehensive feature lists to bias medical diagnosis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 563-572.

Lance, C., LaPointe, J., & Fisicaro, S. (1994). Tests of three causal models of halo rater error. *Organizational Behavior and Human Decision Processes, 57,* 83-96.

Lewis, B., & Linder, D. (1997). Thinking about choking? Attentional processes and paradoxical performance. *Personality and Social Psychology Bulletin, 23,* 937-944.

Lippa, K., Feufel, M., Robinson, F., & Shalin, V. (2016). Navigating the decision space: Shared medical decision making as distributed cognition. *Qualitative Health Research.*

Luchins, A. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological Monographs, 54,* i-95.

MacIntyre, T., Igou, E., Campbell, M., Moran, A., & Matthews, J. (2014). Metacognition and action: A new pathway to understanding social and cognitive aspects of expertise in sport. *Frontiers in Psychology, 5,* 1-12.

Manoharan, T., Muralidharan, C., & Deshmukh, S. (2011). An integrated fuzzy multi-attribute decision-making model for employees' performance appraisal. *The International Journal of Human Resource Management, 22,* 722-745.

McGill, D., van der Vleuten, C., & Clarke, M. (2011). Supervisor assessment of clinical and professional competence of medical trainees: A reliability study using workplace data and a focused analytical literature review. *Adv in Health Sci Educ, 16,* 405-425.

McPherson, S. (2000). Expert-novice differences in planning strategies during collegiate singles tennis competition. *Journal of Sport & Exercise Psychology, 22,* 39-62.

Mitchell, E., Arora, S., Moneta, G., Kret, M., Dargon, P., Landry, G., et al. (2014). A systematic review of assessment of skill acquisition and operative competency in vascular surgical training. *J Vasc Surg, 59,* 1440-1455.

Murphy, K., Jako, R., & Anhalt, R. (1993). Nature and consequences of halo error: A critical analysis. *Journal of Applied Psychology, 78,* 218-225.

Neequaye, S., Aggarwal, R., Herzeele, I., Darzi, A., & Cheshire, N. (2007). Endovascular skills training and assessment. *J Vasc Surg, 46,* 1055-1064.

Nyamsuren, E., & Taatgen, N. (2013). The effect of visual representation style in problem solving: A perspective from cognitive processes. *Plos One, 8,* e80550.

Oh, H., Jo, S., & Myunh, R. (2014). Computational modeling of human performance in multiple monitor environments with ACT-R cognitive architecture. *International Journal of Industrial Ergonomics, 44,* 857-865.

Patel, V., & Groen, C. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science: A Multidisciplinary Journal, 10,* 91-116.

Pennycook, G., Fugelsang, J., & Koehler, D. (2015). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology, 80,* 34-72.

Posner, M. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32,* 3-25.

Protopapas, A., Archonti, A., & Skaloumbakas, C. (2007). Reading ability is negatively related to Stroop interference. *Cognitive Psychology, 54,* 251-282.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavioral and Brain Sciences, 22,* 341-365.

Rasmussen, J., Pejtersen, A., & Goodstein, L. (1994). Cognitive systems engineering. New York: John Wiley & Sons.

Robinson, F. E. (2011). The role of deliberate behavior in expert performance: The acquisition of information gathering strategy in the context of emergency medicine. Unpublished Master's thesis. Wright State University, Dayton, OH.

Sadideen, H., Alvand, A., Saadedden, M., & Kneebone, R. (2013). Surgical experts: Born or made? *International Journal of Surgery, 11,* 773-778.

Schaverien, M. (2010). Development of expertise in surgical training. *Journal of Surgical Education, 67,* 37-43.

Schmidt, R., & White, J. (1972). Evidence for an error detection mechanism in motor skills: A test of Adams' closed-loop theory. *Journal of Motor Behavior, 4,* 143-153.

Schraagen, J. (1993). How experts solve a novel problem in experimental design. *Cognitive Science, 17,* 285-309.

Shah, A., Barto, A., & Fagg, A. (2013). A dual process account of coarticulation in motor skill acquisition. *Journal of Motor Behavior, 45,* 531-549.

Simon, H. (1967). Motivational and emotional controls of cognition. *Psychological Review, 74,* 29-39.

Simon, H. (1973). The structure of ill structured problems. *Artificial Intelligence, 4,* 181-201.

Simon, H., & Gilmartin, K. (1973). A simulation of memory for chess positions. *Cognitive Psychology, 5,* 29-46.

Smith, M., Flach, J., Dittman, S., & Stanard, T. (2001). Monocular optical constraints on collision control. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 395-410.

Smith, D., Greeno, J., & Vitolo, T. (1989). A model of competence for counting. *Cognitive Science, 13,* 183-211.

Solomonson, A., & Lance, C. (1997). Examination of the relationship between true halo and halo error in performance ratings. *Journal of Applied Psychology, 82,* 665-674.

Stanard, T., Flach, J., Smith, M., Warren, R. (2012). Learning to avoid collisions: A functional state space approach. *Ecological Psychology, 24,* 328-360.

Stanujkik, D., Magdalinovic, N., & Jovanovic, R. (2013). A multi-attribute decision making model based on distance from decision maker's preferences. *Informatica, 24,* 103-118.

Tanaka, J., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology, 23,* 457-482

Tien, T., Pucher, P., Sodergren, M., Sriskandarajah, K., Yang, G., & Darzi, A. (2015). Differences in gaze behaviour of expert and junior surgeons performing open inguinal hernia repair. *Surg Endosc, 29,* 405-413.

Timmermans, D. (1993). The impact of task complexity on information use in multi-attribute decision making. *Journal of Behavioral Decision Making, 6,* 95-111.

Toner, J., & Moran, A. (2014). In praise of conscious awareness: A new framework for the investigation of "continuous improvement" in expert athletes. *Frontiers in Psychology, 5,* 1-5.

Tourangeau, R., & Rasinski, K. (1988). Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin, 103,* 299-314.

Tourangeau, R., Singer, E., & Presser, S. (2003). Context effects in attitude surveys: Effects on remote items and impact on predictive validity. *Sociological Methods and Research, 31,* 486-513.

van Hove, P., Tuijthof, G., Verdaasdonk, E., Stassen, L., & Dankelman, J. (2010). Objective assessment of technical surgical skills. *British Journal of Surgery, 97,* 972-987.

Vicente, K.J. (1999). Cognitive work analysis: Toward safe, productive, and healthy computer-based work. Mahwah, NJ: Lawrence Erlbaum Associates.

Vicente, K., & Wang, J. (1998). An ecological theory of expertise effects in memory recall. *Psychological Review, 105,* 33-57.

Weaver, S., Newman-Toker, D., & Rosen, M. (2012). Reducing cognitive skill decay and diagnostic error: Theory-based practices for continuing education in health care. *Journal of Continuing Education in the Health Professions, 32,* 269-278.

Weiss, D., & Shanteau, J. CWS: A user's guide. *Retrieved from https://www.academia.edu/2734414/CWS_A_user_s_guide*

Weiss, D., & Shanteau, J. (2014a). Who's the best? A relativistic view of expertise. *Applied Cognitive Psychology, 28,* 447-457.

Weiss, D., & Shanteau, J. (2014b). Selection effects and the real world. *Applied Cognitive Psychology, 28,* 464.

Wertheimer, M. (1959). Productive thinking. New York, NY: Harper and Row.

Westenberg, M., & Koele, P. (1994). Multi-attribute evaluation processes: Methodological and conceptual issues. *Acta Psychologica, 87,* 65-84.

Wilson, K. (2010). An analysis of bias in supervisor narrative comments in performance appraisal. *Human Relations, 63,* 1903-1933.

Wolfe, J. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review, 1,* 202-238.

Yule, S., Flin, R., Paterson-Brown, S., & Maran, N. (2006). Non-technical skills for surgeons in the operating room: A review of the literature. *Surgery, 139,* 140-149.