Journal of Expertise
2022. Vol. 5(2-3) © 2022.

ISSN 2573-2773

Context Matters in Expert Performance

F. Eric Robinson[1], Valerie L. Shalin[1], Debra Steele-Johnson[1], and Brian Springer[2]

[1]Department of Psychology, Wright State University, USA
[2]Boonshoft School of Medicine, Wright State University, USA

Correspondence: F. Eric Robinson, f.eric.robinson@gmail.com

Abstract

Experts act in dynamic, open environments. In this paper, we combined quantitative and qualitative methods to compare the behavior of resident (trainee, presumably less-expert) and attending (licensed, presumably expert) emergency physician behavior, observed in their natural work environment. Though consistent with insights from both laboratory and observational studies, our analyses allow us to quantify the experts' *constrained* adaptation to variability in the work setting. We examined the effects of patient-, shift-, and system-level variation on expert behavior related to three different abstract task behaviors independent of medical complaint: establishing goals for patients, enacting goals for them, and reducing uncertainty. Our analyses indicate that attending physicians adapted their goal establishment and uncertainty reduction behavior in response to contextual variation, whereas resident physicians did not. In contrast, attending and resident physicians both adjusted their goal enactment behavior, but in response to different contextual features. The comparison of residents' overall lack of context sensitivity with expert behavior indicates that this skill develops subsequent to the formal training period of residency. Our findings underscore the need to account for constrained expert behavioral adaptation across multiple instantiations of apparently similar problems in different contexts.

Keywords

Naturalistic Decision Making/NDM, emergency medicine, medical decision making, situated cognition, expert reasoning

Introduction

Subtle differences in problem context may influence expert behavior, necessitating adjustments in apparently standard approaches. We describe the nature of expert adaptation to context in the domain of emergency medicine. We identify cross-situational adaptability as a crucial element of expertise and suggest domain-general behavioral processes that enable comparisons across a variety of individual patient complaints (for a similar approach, see Donner-Banzhoff et al., 2017). We observed attending and resident physicians in their real-world work context across three different generalized subtasks (goal establishment, goal enactment, and uncertainty reduction), in response to contextual variation in patient difficulty, shift difficulty, and work system. Differences in context sensitivity across residents and attendings indicate that such skill appears to develop after residency is complete. These findings have implications for theory and measurement methods as well as training and assessment.

The remainder of this introduction covers the contemporary conceptualization and measurement of expertise, insights from the development of expertise, the importance of contextual adaptation in medicine, and the insights gained by various methodological approaches to the study of expertise. We conclude the introduction with an overview of the emergency medicine domain examined in the present study, emphasizing that multi-dimensional task content is never ideally sequenced for learning and outcome-oriented feedback is relatively sparse.

Contemporary Conceptualization of Expertise

Two apparently competing methods of operationalizing expertise inform its conceptualization and measurement, contrasting what we call the *outcome approach* and the *process approach*. Ericsson's (2007) criterion of reproducibly superior performance seeks to identify experts based on consistently superior task outcomes. Outcome-oriented accounts assume that desirable outcomes can be operationalized, and that consistently positive outcomes indicate expertise (Ericsson, 2014). Experimental tasks consistent with this approach require a standard/benchmark. Hence, many studies of medical expertise focus on a subtask of medical practice (e.g., diagnosis) in which the problem is pre-specified, and the answer is known to the researcher in advance (e.g., Patel et al., 2013). Such an approach has been particularly productive regarding the analysis of medical error, the prevention of which is the focus of important health care initiatives.

However, relying on outcome as a measurement of expertise across the numerous subtasks in medicine introduces a number of methodological challenges. Most obviously, patient compliance influences outcome. Moreover, in the case of a distributed work system, outcomes cannot necessarily be attributed solely to the actions of the individual professional (Patel, Kaufman, et al., 2014; Vankipuram et al., 2014; Weiss & Shanteau, 2014). Even when comparing outcomes across large samples of patients, it is difficult to control properly for the numerous factors that may influence patient outcomes such as differences in patient population or availability of resources. More subtly, patient outcomes are difficult to operationalize due to their multidimensional nature (e.g., patient satisfaction, symptom relief, quality of life; Catchpole & Alfred, 2018). Not only are these subjective and potentially conflicting, but the weighting of their importance may also differ by patient. Other relevant objectives and constraints include management of medical resources (Ash et al., 2007), regional availability of care facilities, insurance concerns, and billing implications. The combination of goals and constraints renders medical reasoning a *multi-objective decision-making task* (Roe et al., 2001; Roijers, et al., 2013; Voss et al., 1983).

In contrast, Weiss & Shanteau (2014) account for expertise as a process that assures consistent treatment of like problems and different treatment across unlike problems. Because the characterization does not emphasize correctness, but rather process consistency, we call this the *process* approach. The process-oriented account acknowledges that outcomes are not always well defined (Reitman, 1964). Indeed, diagnostic accuracy may be difficult to confirm in the absence of a definitive test. Whereas the outcome approach does not account for the influence of uncontrolled context exerting stochastic influences on outcome (Weiss & Shanteau, 2014), the process approach nicely recognizes the role of socio-culturally determined work practice and better accommodates the uncertainty characteristic of many medical specialties. The primary contribution of Weiss & Shanteau is on the experimental paradigm and mathematical framework to evaluate consistency; they do not specify the cognitive processes that result in the observed consistency. Moreover, the comparison of processes across situations requires judgment of the same controlled set of selected stimuli in an assumed context (Weiss & Shanteau, 2014). Though feasible in a laboratory setting, the Weiss & Shanteau approach lacks theory about how to evaluate problem similarity in real-world conditions where context and content may vary greatly.

Consistent with an emphasis on process, many researchers examine how experts approach problems. For example, some emphasize changes in reasoning processes (Patel & Groen, 1986) arising from the superior framing of a task across various stages of learning (Cheng, 1985; Delaney, 2018; Gobet et al., 2017). Gobet and Chassey (2009) account for expert intuition with template theory, whereby experts assemble chunks of information into schemata-like templates with a core of essential feature classes and additional slots for optional but potentially relevant information (see also Gobet & Simon, 1996). Such models of expert skill emphasize the automatic retrieval and application of static, stored knowledge structures (Tenison & Anderson, 2016) assuming problem typicality. Consistent with a process-oriented framing of Naturalistic Decision Making (NDM), experts may not deliberate or compare options. Instead, they act rapidly and intuitively based on their evaluation of a given situation and associated key cues. A series of papers postulate that, for the highly proficient expert, a recognition-based process of rapidly evaluating situations results from stored patterns gained through experience (Klein, 2015; Kahneman & Klein, 2009). Catchpole and Alfred (2018) and Falzer (2018) suggest that the NDM perspective would better capture human expertise in medical decision making, particularly for accommodating context.

Development of Expertise

A classic psychological approach to the analysis of expertise is to examine the development process, generally by comparing different cohorts at different levels of expertise. Using qualitative data, Dreyfus and Dreyfus (2005) described a five-stage model of developing expertise in which the learner gradually moves from using context-free rules to operating based on context-driven perception of the problem and the associated solution. The question here is not so much the final state, but the sequential learning process whereby that occurs. Benner (2004) describes the application of this approach to the study of nursing expertise.

Experience alone is insufficient to generate expertise (Kahneman & Klein, 2009; Phillips et al., 2004). The structure of the environment and opportunity to learn relevant environmental features help determine whether experience leads to expert intuition (Kahneman & Klein, 2009; Klein, 2015). Experts may also enrich their own experiences through self-reflection to gain new insight or learn from past events (Klein, 1998, cited in Phillips et al., 2004). Ericsson (2006; 2008) proposed that expertise develops via an intensive process of deliberate practice that targets specific skills, initially under the guidance of a coach but later directed independently.

Feedback also plays an essential role in most accounts of human learning. Yet, compelling analysis of language or category acquisition (Lecun, 2017; Putnam, 1979; Quine, 2008; Vong et al., 2016) and computational models of learning (Jordan & Mitchell, 2015) suggest that the multi-dimensional complexity of open-world learning problems and the sparsity of feedback under-determine the desired mapping between context and response. The absence of feedback creates a more time-consuming and computationally complex learning problem (Jordan & Mitchell, 2015).

Cognitive accounts of learning in the experimental literature often assume carefully constructed formal training of generalizable content, presumably for reasons of efficiency (Van Lehn, 1987). Lessons must be carefully scoped and sequenced. Sweller and Chandler (1994) suggested that skills are learned most effectively when schemata can develop gradually without overburdening cognitive resources at any one time. However, researchers have acknowledged the supportive role of context and content in the acquisition of general principles for some time (Cognition & Technology Group at Vanderbilt, 1993), a view closely aligned with situated learning (Brown et al., 1989) and the benefits of apprenticeship (Lave & Wenger, 1991). Perhaps paradoxically, contextual detail facilitates rather than impedes learning. In fact, contextualized, problem-based learning is widely applied in medical training (Ibrahim et al., 2018; Preeti et al., 2013; Servant-Miklos, 2018), although scenario quality contributes to the method's success (Azer, 2007; Azer et al., 2012). In this paper we raise additional contextual dimensions that might be explored earlier in structured medical training.

Contextual Adaptation in Medical Expertise

The present study examined differences in context-sensitivity between emergency medicine residents and attending physicians. Well-intentioned efforts to standardize medical intervention aim to reduce outcome variability (e.g., MERIT Study Investigators, 2005), but can also falter. This occurs because although the patient's complaint or condition may be defined, the specifics of the problem to be solved may vary. We are certainly not the first to investigate the influence of contextual features on the care process (Alqahtani et al., 2018; Schmidt et al., 2017). Indeed, a prominent criticism of current trends in medicine toward adherence to guidelines is that they oversimplify complex situations and neglect the role of expertise in treating individual patients (Catchpole & Alfred, 2018; Falzer, 2018). The macrocognitive literature (Patterson et al., 2010) clearly acknowledges the importance of environmental features such as changes over time, complex causal relationships between variables, conflicting goals, or workload. In recognition of such influences Klein (2007a; 2007b) coined the term *flexecution* as characteristic of expert performance.

The adaptive expertise literature focuses on the qualities of those who are able to adjust plans, goals, or procedures in response to developments in the world (Carbonell et al., 2014; Ward et al., 2018). However, the corresponding literature focuses on the evolving context of a single problem over time. We argue that it is equally important to account for ways in which experts navigate (even mundane) variability between multiple instantiations of similar problems. No two problems are ever exactly alike, blurring the distinction between routine and novel problems (Robinson et al., 2020). Consistent with Klein's (1989) recognition-primed decision making and Vankipuram et al. (2014) specifically in medicine, we provide quantitative evidence that experts are better able than novices to approach different instances of similar problems in different ways depending on situational requirements. This supports our claim that expertise is revealed not just by the ability to adapt within a single problem over time, but by the ability to adapt across multiple contexts.

Because of the requirements for professional certification and predictability in distributed work (Shalin et al.,1997), we emphasize *constrained flexibility* in reasoning. Acknowledging constraints accommodates the role of disciplinary guidelines with case specifics. Expertise is executed in a physical environment that includes other people and resources that influence reasoning (Hutchins, 2005; Zhang & Norman, 1995). However, just as action is constrained in technical (Rasmussen et al., 1994; Vicente, 1999), and intentional (Wong & Kodagoda, 2015; Wong et al., 1995) systems, a larger culture of medical practice also imposes constraints on patient care (Myneni et al., 2014).

Methodological Implications

Our methods diverge from both conventional psychological experimentation and the particular qualitative methods characteristic of Naturalistic Decision Making. Performance on controlled experimental tasks is the foundation of cognitive psychology for a number of very good reasons. These include the attribution of causality to theoretically significant predictors, the ability to collect specialized performance measures that reveal cognitive processes, and the use of quantitative analysis to determine the statistical significance of the influence of predictors on behavior. Comparison across conditions (including levels of expertise or, as we examine here, hospital) is critical. Ideally, a corpus of theoretically inspired experimental studies eventually adds up to a complete picture of the capability in question. Nevertheless, researchers in medical cognition recognize the potential influence of the selected research method on findings (Patel, Kaufman, et al., 2014).

To emphasize our contribution, we briefly clarify differences in established alternative qualitative methods (Creswell, 2007). In narrative research, researchers gather stories from an individual or a few individuals and then analyze them for key elements. The critical incident method associated with NDM (Crandall, Klein, & Hoffman, 2006) best fits this category. In phenomenological research (e.g., Benner, 2004), researchers interview subjects regarding their experience of events. Event-prompted narratives as in Mogford et al. (1997) or Hoffman et al. (2006) are somewhat consistent with this method. Researchers conducting ethnographic studies gather data from interviews and observation to provide a coherent understanding of the culture's norms, values, and constraints (Forsythe, 1993). Conventional cognitive task analysis (Schraagen et al., 2000) has the potential to fit this category, though typically focuses more on procedurally oriented cognitive accounts of particular domain subtasks (e.g., running test processes in a manufacturing plant; Gore et al., 2018). Influenced by an informal sampling of context (see Dubois & Shalin, 1995 for an exception), these qualitative approaches cannot argue quantitatively for the importance of different variables. Case studies examine one or a few specific cases of a phenomenon in-depth by gathering information from multiple sources. Some of Reason's (1990) work on accident/error analysis may fit in this category. Using a grounded theory method as in the present paper, the primary issue is the development of the coding scheme itself. Researchers will gather information regarding the phenomenon they are interested in, and then annotate that information in an open, emergent coding scheme. The observer has substantial discretion over the categories. For example, Franklin et al. (2014) view observations through a decision-making lens and Payne & Patel (2014) focus on heuristics and biases associated with diagnosis. These analytic choices cannot be wrong, but they are different. In grounded theory, sufficiency is evaluated with respect to saturation criteria, meaning that additional observations do not alter the conclusions. Triangulation, confirming the same conclusions across methods or context, assures generality.

Taking inspiration from experimental design, the nested (hierarchical, multilevel) analysis of observational data partitions the effect of context and correlated predictors. Here, patient cases are nested inside doctors who are associated with specific shifts and hospitals. In this way, we can examine doctor responses to patient cases *in situ* using different measures of behavior, providing general conclusions about how doctors cope with variability associated with the work setting (shift demands and hospital) separate from variability between specific cases.

One of the challenges of quantifying behavior in the naturally occurring medical domain is finding an appropriate level of abstraction. Overly specific characterizations will fail to provide a sufficient number of similar instances to reveal patterns. Highly abstract characterizations will almost certainly cross established disciplinary distinctions; e.g., the different specific orders associated with managing a cardiac complaint relative to a gastrointestinal complaint. To address the behavioral abstraction problem, Robinson et al. (2020) used data-driven factor analysis to identify three primary behavioral measures exhibited by emergency physicians to evaluate and care for their patients, each representing a different aspect of the care process independent of the particular patient complaint. Rather than focus on a subtask of care, these measures collectively encompass a broad sample of tasks associated with care delivery in the emergency department (ED).

Goal establishment behaviors rely on the patient interview to allow the doctor to navigate an open-ended problem space and determine what should be done for a patient. We distinguish a lower level of *goal enactment* behaviors that allow the physician to implement the care plan in the broader sociocultural context of the work system. Consistent with *flexecution* (Klein, 2007a; 2007b), goal establishment and goal enactment interact with one another in an iterative process. Within a case, patient interactions and test results can drive new questions and tests, and physicians may adapt their goals in response. Nevertheless, our focus here is on the adaptation revealed between cases. Physicians use a third process, *uncertainty reduction*, to confirm aspects of technical knowledge such as medication dosage. Note that these constructs are independent of complaint, equally pertinent to heart conditions, drug abuse, and injuries. Though emergent from data, the distinction between goal establishment and goal enactment is consistent with the action hierarchy of Vallacher and Wegner (2012) and the work domain ontology syntax in Franklin et al. (2014) and Norman (1986). The distinction is also embedded in classic problem-solving architectures (Anderson, 1983; Laird et al., 1987), and consistent with distinctions made between needs assessment and problem response in medical decision making (Payne & Patel, 2014). To wit, determining *what* to do is different from determining *how* to do it.

The Practice of Emergency Medicine

Our study occurs in the domain of emergency medicine. Emergency medicine is a dynamic, time-sensitive domain in which physicians must simultaneously address multiple patient complaints of varying severity. In addition, high patient volume and changing patient conditions generate intense time pressure. Emergency physicians' primary goal is to evaluate and treat emergency medical conditions. As part of that goal, emergency physicians must determine whether a patient should be discharged to follow up care or should be admitted to the hospital. As such, emergency physicians rarely receive definitive feedback on a patient's ultimate diagnosis or outcome. Relatively infrequent and somewhat distal feedback about care processes may be obtained from patients who return after discharge, pushback from floor specialists who must agree to receive admitted patients, patient satisfaction surveys, interactions with staff members (e.g., negotiating with the charge nurse), or surprising laboratory results not expected based on the interview.

More broadly, emergency medicine is a culturally constrained domain with values and norms guiding effective practice. Emergency physicians in the United States work within a complex network of care providers, each with unique perspectives and priorities. Doctors must be able to justify their decisions to patients and the larger medical community and must work with others to enact their plans (Johnson et al., 2008; Moran-Barrios & Gauna-Bahillo, 2010). Adherence to norms constrains response and facilitates coordination among the community of health providers (Croker et al., 2008; Foy et al., 2010; Johnson et al., 2008; Martin et al., 2010). Such norms and processes help to ensure generally acceptable, justifiable outcomes despite shifting contextual constraints and in the absence of individual feedback. Despite this, physicians can alter their work style to some extent to suit their individual or patients' needs.

Emergency physicians who are licensed to practice independently are known as attending physicians (also referred to here as "attendings"); prospective emergency physicians who are completing their specialty training after medical school are known as residents. During residency, physicians-in-training receive didactic training as well as hands-on training via practice oral board cases and high-fidelity simulation. In addition, residents work in real-world clinical conditions under the guidance of a more experienced attending. In contrast to formal didactic training, the clinical environment is not controlled for cognitive demand or combinations of relevant case features. Residents are supervised, but are not necessarily observed; for instance, they will often conduct patient interviews without the attending physician present. Resident physicians must learn to adapt safely and efficiently to real-world variation in patient characteristics, workload, and work systems, without many of the immediate and formal feedback mechanisms normally found in more structured training. This paper identifies the acquisition of such contextual adaptation skill as a key feature of developing expertise.

The Present Study

Variation in work practice and context can affect decision making in the ED (Franklin et al., 2014). We build on a prior data set (Robinson et al., 2020) to provide quantitative evidence that one of the features that distinguishes experts from novices is the ability to adapt behavior to variations in work context. Relative to Robinson et al. (2020), we added a larger sample of attending physicians to the original data set to facilitate direct comparisons of resident and attending physicians and thereby demonstrate saturation in the original coding scheme. We derived criterion measures based on the observed physician behaviors, then utilized multilevel modeling of those derived measures to describe changes associated with variation in patient-, shift-, and hospital-level variables. Quantitative results were supplemented by qualitative examples. This analysis allowed us to identify the adaptive capability that experienced attending physicians have mastered that does not appear to be accounted for during the formal training period of residency. Differences between the most experienced doctors and their less experienced counterparts provided insight into the evolution of context sensitivity. We hypothesized that attendings would be more sensitive than residents to contextual features, evidenced by adjustments to goal establishment, goal enactment, and uncertainty reduction associated with variation in our predictors.

Method

The methods and measures described below are the same as those used in Robinson et al. (2020). This study was approved by the institutional review boards (IRBs) at the hospitals under observation, as well as the university with which the observed hospitals were affiliated. We expected the distribution of patient complaints to differ across time (i.e., the distribution of complaints late Saturday night is likely to differ from the distribution of complaints early Wednesday morning). We thus attempted to control for such temporal confounds by counterbalancing observations across hospital, part of the week, and time of day.

Subjects

We recruited qualified emergency physicians beyond the first year of residency who worked in the EDs under observation, identified based on compatibility between their work schedule and the aforementioned counterbalancing scheme. The residents in this sample worked in several local hospitals, rotating periodically. Attending physicians in our sample typically worked in only one or two hospitals. We recruited participants via individual emails, general announcements at weekly lectures attended by residents as part of their training, and staff meetings. We combined the original data from Robinson et al. (2020) with new observations not previously reported. The prior data set contained observations from nine second-year residents, eight third-year residents, and nine attending physicians. We conducted additional observations to increase our sample size (adding 11 attending physicians and one third-year resident). The final data set reported here therefore contained 38 physicians (nine second-year residents, nine third-year residents, and 20 attending physicians). After excluding outliers and patients who could not give meaningful consent to being observed (see below), the observed physicians treated a total of 331 individual patients included in the analysis (147 patients seen by residents and 184 patients seen by attending physicians).

Observation Setting

Observations occurred over a period of approximately two years in the EDs of two different teaching hospitals associated with the medical school of a public university in the midwestern United States. One hospital was a suburban ED with approximately 40 beds that served patients that were primarily older, Caucasian, and insured. The other hospital was an urban ED with approximately 60 beds that served patients of varying age, with a large percentage of minority and uninsured patients relative to the suburban ED. At the time of observation, the suburban hospital used a partially paper-based record system including T-sheets™; these problem-specific forms served to document the patient interview based on predetermined fields. The urban hospital used a primarily electronic records system. Reflecting higher patient volume, nurses at the urban hospital had a greater tendency to order lab tests in triage (prior to a patient being seen by a physician) compared to nurses at the suburban hospital as part of an effort to reduce overall visit time and increase patient throughput.

Procedure

Each "observation" focused on a single physician as they cared for multiple patients over the duration of one full work shift, usually 10 hours or longer. Data were recorded and coded at the patient level, within each observation (i.e., patients were nested within doctors). A single observer shadowed each physician one time as they performed their work tasks, including patient interviews and exams, documentation, and interacting with hospital staff. The same observer conducted all observations. The observer took handwritten notes of all directly observable actions or thoughts articulated by the physician, either to the researcher or in conversation with staff members. A stopwatch was used for timestamps. Information that could be used to identify patients or doctors was not recorded at any time; patients were assigned an identifier (e.g., patient 1) used to identify notes associated with their care while in the ED.

Unobtrusive observation was not possible in this setting; the observer therefore adopted a level of participation between unobtrusive and participant observation. The observer asked questions for clarification as well as general questions to avoid potentially leading the doctor (e.g., "what are you thinking about this patient?" or "do any lab values jump at you?"), being careful not to disrupt ongoing activities (e.g., asking while walking between rooms). Doctors sometimes asked the observer to get a blanket or something similar; this was documented and coded as if the doctor had asked a staff member to do the same or had done it themselves.

Patients gave verbal consent for the observation of their care after receiving an explanation of the study from the doctor who was being shadowed. The observer left the room during sensitive exams such as rectal or pelvic exams. All patients who were asked agreed to allow the observer to be in the room for interviews, but in some cases the doctor forgot to explain the study to the patient, or the patient was not in a state to give consent (e.g., unconscious or incoherent). Data from these patients were excluded from the analysis.

Analysis

Both predictor and criterion measures were derived from the observational data. Predictors reflected situational or contextual factors that influenced physician behavior. Calculated criterion measures reflected counts of specific doctor behaviors coded in the observation notes. We describe the predictor variables first, followed by a description of the low-level behaviors that are reflected in the criterion measures. Our criterion measures were calculated based on the results of an exploratory factor analysis; criterion measures are therefore described in the results section.

Predictor Variables

We used differences between patients and doctors to predict physician behaviors. We describe each predictor variable in turn.

Patient Level Predictors. We used a measure of *patient difficulty*, which served to capture the effect of individual patients on physician behavior as well as help operationalize the overall workload of a shift (see below). Patient difficulty ratings were generated by the observer during data analysis based on a subjective assessment of the time and effort each patient required of the doctor, relative to the general population of patients. Patient difficulty scores reflected more than medical complexity; the assessment included traditional components such as the number of tests or interventions performed, but also accounted for the time spent consulting with other physicians, how atypical or complex the physician judged the patient's case to be, effort required to accommodate patient-specific factors such as comorbidities or access to follow-up care, how cooperative or demanding the patient was, etc. These captured aspects of the doctor's workload overlooked by other measures such as urgency-focused triage scores. We used a three-level rating:

- Patients who required minimal workup or intervention (commonly referred to in the

ED as "treat and street" patients) were scored as "1." Examples of patients who often (though not always) received a 1 would be those seeking treatment for sinus infections, minor cuts needing stitches, etc.
- "Average" patients were scored as "2"; e.g., a textbook presentation of shortness of breath or chest pain after exertion in a patient with a history of heart problems.
- Difficult/uncooperative patients or those with complicated medical issues who required more time or attention compared to most patients were scored as "3."

Doctor Level Predictors. We used three predictors to describe differences between physicians: the difficulty of the doctor's shift, years of clinical experience, and the hospital in which the doctor worked.

- Shift Difficulty. Whereas patient difficulty reflected the influence of workload at the level of an individual patient, *shift difficulty* captured workload across the entirety of a shift. Patient care is the physicians' primary responsibility during a shift; patient difficulty scores therefore formed the basis for shift difficulty. Shift difficulty was calculated as the sum of the patient difficulty scores of the patients under the doctor's direct care during a shift (i.e., patients seen primarily by a resident under an attending physician's supervision did not count towards the attending's shift difficulty score). Shift difficulty scores ranged from 9 to 44.
- Experience. We analyzed attendings and residents as separate groups and explored potential effects of experience within each cohort (e.g., whether more experienced attendings behaved differently than less experienced attendings). We operationalized experience according to the number of years that the physician had been in clinical practice. Second- and third-year residents were considered to have two and three years of experience, respectively. The attending physicians had between five and 34 years of experience, with a mean of approximately 15 years.
- Hospital. The two hospitals in which our observations occurred differed by work system (i.e., medical records system) as well as work practices (e.g., whether nurses were likely to order labs in triage). We therefore included hospital as a nominal variable.

Criterion Measures

Our criterion measures reflected the observable behavior of the shadowed physicians. We identified behaviors of interest *a priori* driven by our research questions and literature review, refined during observation of weekly lectures over approximately six months. We further refined our coding scheme based on testing using pilot observation notes, but the coding scheme was finalized before being applied to any observations reported here.

We coded different components of patient care and disposition (behaviors) in the ED (Table 1; see the Appendix for the full coding manual), according to the following categories:

- *Information gathering* behaviors helped the doctors gather facts about the patient's case. These behaviors were further distinguished by the type of information gathered and its source.
- *Diagnostic* behaviors helped the doctors identify or eliminate causal hypotheses for symptoms.
- E*vidence evaluation* behaviors helped the doctors evaluate the completeness of their understanding or the quality of available evidence.
- *Patient management* behaviors helped the doctors treat or manage patients during their stay in the ED.
- *System management* behaviors helped the physicians to work within the constraints imposed by the hospital or broader healthcare system.
- *Filtering* behaviors limited the problems to address or identified the scope of the patient's complaint.

Table 1. Coding Categories Used in the Present Study (Robinson et al., 2020; reused with permission)

Behavior Category	Subcategory
Information gathering	
Type	
	Current symptoms
	Timeline
	Past medical history
	Contributors
	Reference information
	Other
Source	
	Exams
	Patient
	Tests/images
	Patient's family/friends
	Medical records
	Hospital staff
	Internet/reference materia
	Miscellaneous
Diagnostic behavior	
Evidence evaluation	
Patient management	
	Treatment
	Consulting
	Collaboration
	Logistics
System management	
Filtering behavior	

Note. Some behavioral categories were further divided into subcategories to identify actions of interest more specifically. Italicized behaviors represent the actual codes assigned to data. Each behavior was assigned to a single category with the exception of information gathering, for which both a single type and a single source were assigned. For example, the patient stating to the doctor that they have a headache would be coded as obtaining information about "current symptoms" (type) from "the patient" (source).

Our analysis followed the same overall process as used in Robinson et al. (2020). Observation notes were transcribed into a spreadsheet, and then individual behaviors were categorized according to the above coding scheme. Individual behaviors were identified as discrete actions (e.g., getting a blanket) or as distinct pieces of information (e.g., individual questions or statements). In the event that multiple questions or statements were given at once, the utterance was broken up into items that could reasonably stand alone (e.g., "Do you have a family history of heart problems or cancer?" would be coded as two separate items - "Do you have a family history of heart problems?" and "Do you have a family history of cancer?"). A single code was assigned to each behavior, with the exception of information gathering behaviors (for which both type of information and source of information were coded).

The number of instances of each coding category were summed within each patient for each doctor, allowing us to examine differences in behaviors associated with patient-level predictors (patient difficulty) and doctor-level predictors (experience, shift difficulty, and hospital). We used factor analysis to group the various coded behaviors into interpretable categories representing higher level constructs. Factor scores based on the frequency of behaviors within each factor served as criterion measures in later analyses.

Reliability Check

The first author served as the coder for all observational notes. Several months later, the same coder re-coded a sample of nine observations to assess code-recode reliability (six observations re-coded as part of the reliability check for Robinson et al., 2020, plus an additional three observations).

Patient Difficulty. We assessed the reliability of the patient difficulty ratings by reassessing all of the patients (n = 84) from the nine observations used for the broader reliability check. Weighted Cohen's Kappa was 0.70, indicating reliable ratings. As described above, the "average" patient was supposed to be scored as a two. The mean patient difficulty rating across all 331 analyzed patients was 1.97, with a distribution of 72 patients receiving a "1," 198 receiving a "2", and 61 receiving a "3." The appropriate mean and symmetrical distribution imply sound scoring.

Behavioral Coding. We calculated the number of instances in which a specific behavior was positively identified in both coding sessions, negatively coded (i.e., identified as absent) in both coding sessions, and instances of disagreement between sessions (i.e., coded as present in one but absent in the other). These categories were summed across all of the recoded shifts to ensure adequate sample size. Reliability for each behavior was assessed individually using Cohen's Kappa. Kappa values ranged from 0.47 to 0.93; the least reliable variables eventually fell out of the analysis (described below) and ultimately did not contribute to the findings of this study. Kappa values for the variables included in the final analysis ranged from 0.66 to 0.90.

Results

We first conducted an exploratory factor analysis to aggregate the individually coded doctor behaviors into conceptually related groups of care-related behaviors ("factors") independent of complaint. We analyzed the effects of context on each of these factors separately, using a combination of multilevel modeling and regression to identify differences in how attending and resident physicians adjusted to variation in patient characteristics, workload, and work setting. These quantitative analyses are supplemented with qualitative examples to illustrate the possible nature of any identified differences.

Factor Analysis for Data Reduction

Analysis Procedure

Ten outlier patients (exceeding +/- 5 standard deviations from the mean on any single behavior) were excluded from the analysis, resulting in a final sample of 331 patients after accounting for other exclusions discussed above (147 treated by residents, 184 treated by attendings).

We examined each of the coded behaviors in an iterative factor analysis. Standard guidelines for identifying factors recommend a loading of at least 0.3 on a given factor, with at least a 0.3 difference between loadings on multiple factors (Gorsuch, 1974). We adopted a more conservative approach due to the imprecision in our data set, retaining a variable (i.e., a doctor behavior) if the loading exceeded 0.5 on a single factor and the difference in that variable's loading on multiple factors exceeded 0.34. We then re-ran the factor analysis using only the variables that met the above criteria, iterating until all remaining variables met the loading criteria (three iterations). A three-factor model ultimately survived examination after the one- and two-factor models failed to yield conceptually coherent and complete factors.

Identified Factors

Factor 1 in the final analysis contained the *evidence evaluation behaviors, diagnostic behaviors, system management behaviors, using tests and images to gather information, and logistic behavior*, hereafter referred to as the "goal enactment factor." In this context, labs primarily serve to facilitate or confirm admission goals rather than to form goals (Patel et al., 1997).

Factor 2 contained using the *internet or reference material* to find *reference information* hereafter referred to as the "uncertainty reduction factor."

Factor 3 included finding out about *contributing factors* and *using the patient as a source of information*. Because such behavior occurs primarily at the outset of the patient's visit, we refer to this as the "goal establishment" factor. The final contents of the factors replicated the results of the factor analysis in Robinson et al. (2020). The unrotated eigenvalues for the final three factors were 3.11, 1.92, and 1.36, respectively.

Factors as Dependent Measures

Each construct (goal establishment, goal enactment, or uncertainty reduction) represents a higher-order aspect of the care process. Factor analysis supported the aggregation of low-level variables into factor scores. At the patient level, each low-level variable within a factor was converted into a Z score and then averaged together with the Z scores of the other variables in a factor to obtain an aggregated factor score. For example, to generate a factor score for the goal establishment factor, the total number of coded instances of *contributing factors* and *using the patient as a source of information* were each calculated for a given patient. Those totals were converted to separate Z scores, then averaged together within patients to generate a factor score associated with each patient. The averaged Z scores represent how extreme an individual doctor's behavior was for that construct when treating a given patient, relative to the entire sample of 331 patients. Factor scores served as dependent measures in the remaining analyses.

Multilevel Modeling Overview

We used multilevel modeling to analyze the effect of contextual features (patient difficulty, shift difficulty, experience, and hospital) on factor scores within the attending and resident physicians. Multilevel analysis allowed us to account for the nesting of patients within doctors and the resulting non-independence of our observations (e.g., Brooks et al., 1991). Variance in physician behavior is explained by predictors varying across individual patients (Level 1; patient difficulty, in this case) and across doctors (Level 2; experience, shift difficulty, and hospital). Multilevel modeling examines variance in both intercept (conceptually similar to a main effect in an ANOVA) and slope (an interaction), in separate sequential steps. Intercept variance represents the mean differences in factor scores associated with a given predictor, whereas slope variance represents differences in the relationship between the Level 1 (patient-level) predictors and factor scores across Level 2 (doctor-level) predictors (e.g., whether the relationship between patient difficulty and goal establishment differs across hospital). In our study, significant effects of patient difficulty indicate within-doctor adaptation across different patients, whereas effects of hospital, shift difficulty, and experience reflect between-doctor differences (i.e., doctors who worked in one hospital behaved differently than doctors in the other hospital). Differences in the effects of our predictors across residents and attending physicians reflect differential sensitivity to these contextual features across the two levels of expertise.

Our analyses followed the model-building steps recommended by Bliese (2002), with the exception that we explored slope variance regardless of the result of the deviance chi-square test due to the low power of the test (LaHuis & Ferguson, 2009). In the event that the use of multilevel modeling was not supported (indicated by an ICC < 0.10, demonstrating that the effect of patient nesting within doctors was negligible), we used multiple regression. An alpha level of 0.05 was adopted for all analyses. The assumption of normality in the residuals was violated (particularly for uncertainty reduction behaviors). Although this does not necessarily negate our findings (Knief & Forstmeier, 2020), our results are best interpreted as exploratory to help guide further investigation of expert reasoning.

Contextual Influences Associated with Factor Scores

Goal Establishment

We first examined the effects of contextual features on goal establishment processes, examining the goal establishment behaviors of residents and attending physicians using separate multilevel analyses. We found that attending physicians adjusted their goal establishment processes in response to different contextual factors, but residents did not. We illustrate our quantitative results with qualitative examples.

Quantitative Analysis of Goal Establishment

Residents

Residents' goal establishment factor scores had a mean of 0.28 ± 0.93. The intraclass correlation coefficient (ICC) for the goal establishment factor for the 18 residents was 0.36, indicating substantial between-doctor variance on this factor. However, none of our predictors accounted for variance in resident goal establishment behaviors at any level of analysis, indicating that the residents did not seem to adjust their goal establishment behaviors in response to the contextual features studied here (see Table 2).

Table 2. Results for Residents on the Goal Establishment Factor

Model and Parameter	Parameter Estimate	*SE*	*t*	*df*	*p*
Level 1					
Patient difficulty	0.11	0.11	1.00	142	0.32
Level 2 (Intercept)					
Experience	-0.12	0.31	-0.39	14	0.70
Shift difficulty	-0.02	0.03	-0.84	14	0.42
Hospital	-0.24	0.31	-0.77	14	0.46
Level 2 (Slope)					
Experience	0.07	0.22	0.31	14	0.77
Shift difficulty	-0.03	0.02	-1.68	14	0.12
Hospital	0.35	0.22	1.62	14	0.13

Attendings

The mean goal establishment score for the attending physicians was -0.26 ± 0.66. The ICC for the goal establishment factor for the 20 attending physicians was 0.24, indicating substantial between-doctor variance for attending physicians on this factor. Unlike the residents, Level 1 analyses indicated patient difficulty was related to goal establishment behavior such that individual attending physicians demonstrated more goal establishment behaviors with more difficult patients (Table 3). Level 2 analyses indicated that shift difficulty and the hospital where the doctor worked predicted variance in the intercept for the attending physicians. Unlike the residents, attending physicians who worked at the suburban hospital demonstrated more goal establishment behaviors than attending physicians at the urban hospital, and attending physicians who worked during busy shifts demonstrated fewer goal establishment behaviors than attending physicians with slower shifts. We used Level 2 predictors to investigate slope (interaction) variance but were unable to account for differences in the observed relationship between patient difficulty and goal establishment behaviors with any of our predictors.

Table 3. Results for Attending Physicians on the Goal Establishment Factor

Model and Parameter	Parameter Estimate	*SE*	*t*	*df*	*p*
Level 1					
Patient difficulty	0.26	0.07	3.77	179	< 0.01
Level 2 (Intercept)					
Experience	0.00	0.01	0.02	16	0.99
Shift difficulty	-0.02	0.01	-2.79	16	0.01
Hospital	0.48	0.14	3.54	16	< 0.01
Level 2 (Slope)					
Experience	0.01	0.01	0.73	16	0.48
Shift difficulty	-0.01	0.01	-0.76	16	0.46
Hospital	0.30	0.16	1.84	16	0.08

Note. Predictors in **bold** are significant at $p < 0.05$.

Qualitative Analysis of Goal Establishment

Quantitative analyses suggest that the attending physicians are more responsive to their environment than the residents. Whereas the attending physicians adapted their patient interviews based on the patient, the hospital, and their shift workload, the residents' behavior did not vary systematically. The following example illustrates the importance of goal establishment to the care process, and the influence of patient characteristics in decision making.

> *A patient with a history of hypertension is in the emergency room with a severe headache. After getting the patient's history and performing the exam, the resident reports to the attending physician and proposes to give Lopressor (a blood pressure medication) to lower the patient's blood pressure. The attending says that is fine as long as the patient isn't on cocaine. The resident had not asked about cocaine use and returns to the room to check; the patient reports using cocaine 3 days ago. The resident had to use a different blood pressure medicine instead. When asked by the observer why the patient couldn't get the first medicine, the resident says that with the cocaine it could have sent the patient's blood pressure and pulse higher because "beta" would be blocked by the medicine and "alpha" would be stimulated by the cocaine.*

The resident failed to inquire about a key part of the patient's history during the goal establishment phase, with implications for later goal enactment decisions such as what medication to use. This is also a relatively rare instance of feedback about goal establishment behavior in a real-world clinical setting. The two different interviews in Table 4 illustrate potential differences between resident and attending physicians on goal establishment.

Table 4. Contrasting patient interviews from a resident and attending physician. Both patients presented to the ED with a complaint of chest pain after doing chores. Both physicians had higher than average goal establishment scores within their experience cohort. The attending physician worked at the suburban hospital and the resident worked at the urban hospital. The resident's shift was a difficulty of 20, and the attending physician's shift was a difficulty of 21. Sections of the interviews have been removed to improve readability, noted within the text of the figure.

	Attending Physician Example		Resident Physician Example	
Line	Speaker	Utterance	Speaker	Utterance
1	Attending	What brings you to the ED?	Resident	Why are you here on this lovely day?
2	Patient	I had crushing pain across my chest suddenly.	Patient	I guess I just don't feel good.
3	Attending	What time did that happen?	Resident	How did this start?
4	Patient	About an hour and a half ago.	Patient	I was carrying coal. I guess it was about 300 pounds total. I felt fine, though.
5	Attending	How long did it last?	Resident	Is that a normal activity for you?
6	Patient	About 10 minutes.	Patient	Yes.
7	Attending	What were you doing when it happened?	Resident	When was this?
8	Patient	I was doing yard work and then I sat down, and it started.	Patient	Around 7. Later I had a dizzy spell, black and dizzy.
9	Attending	Did you feel short of breath or sweaty or pale?	Resident	When did that happen?
10	Patient	No, none of that.	Patient	Around 8. I sat on the couch and felt better but then I felt some pressure and some pain. It wasn't too bad; it wasn't sharp, but it was different than the pressure.
11	Attending	Did you get nauseous or vomit?	Resident	How long did it last?
12	Patient	No.	Patient	I had pain off and on for about 30 seconds each.
13	Attending	And the pressure goes across both sides of the chest?	Resident	Have you felt the pain since then?
14	Patient	That's right.	Patient	No, just the pressure.
15	Attending	Does the pain go into your jaw, neck, or back?	Resident	Are you nauseous?
16	Patient	No. I got x-rays on my back two weeks ago for back pain, but they were ok.	Patient	No.
17	Attending	What was the back pain?	Resident	Are you short of breath?
18	Patient	They suspected something with my sciatic nerve.	Patient	Not now but off and on for the last two weeks.
19	Attending	Have you had any fever or cough?		[23 lines removed for space.]
20	Patient	No.	Resident	Try to relax. (Smiling) It's easy for me to say… Do you have any family history of heart problems before age 60?
21	Attending	Does anything help or make the pain worse?	Patient	No. My sister had cancer, though.
22	Patient	No. I took aspirin when it happened, too.	Resident	Do you have any personal history of anything like that?
23	Attending	Is there any discomfort now?	Patient	No.
24	Patient	No.	Resident	Do you have diabetes?
25	Attending	Have you noticed any swelling in your legs?	Patient	No.
26	Patient	No.	Resident	Do you have high blood pressure?
27	Attending	Is there any abdominal pain?	Patient	Yes.
28	Patient	No.	Resident	What meds are you on?
29	Attending	Are you taking any meds for any other conditions like diabetes?	Patient	(The patient tells him the names and doses, but I can't catch them.)
30	Patient	(The patient tells the attending what medications she uses but I miss them.)	Resident	How long have you been on those?
31	Attending	Are there any surgeries I should know about?	Patient	Years.
32	Patient	No.	Resident	Have there been any recent changes to your meds?
33	Attending	Have you ever had a heart attack or a stroke?	Patient	I'm off (I can't hear the name) for the stress test.

Table 4 continued on next page.

Table 4, continuation

	Attending Physician Example		Resident Physician Example	
Line	Speaker	Utterance	Speaker	Utterance
34	Patient	No.	Resident	Have you had any recent travel?
35	Attending	Have you ever had a clot?	Patient	No.
36	Patient	No.	Resident	Have you had any broken bones?
37	Attending	Have you ever had a stress test?	Patient	No. I had a hernia repair three years ago, though.
38	Patient	No.	Resident	Any other surgeries?
39	Attending	Are there any other conditions that you want me to know about?	Patient	No.
40	Patient	I had a lumpectomy years ago.	Resident	[Content removed for space]...How is the pain now from 1-10?
41	Attending	Are you a smoker of have you ever smoked?	Patient	The pressure is a one or a two.
42	Patient	No. I have a history of stroke in my family, though.	Resident	How was the last pain?
43	Attending	Have you been taking all of your blood pressure meds?	Patient	Maybe a three.
44	Attending	Can you show me where the earlier back pain was? [Two lines removed for space.]	Resident	(Resident pushes the patient's chest.) Does it hurt to push on your chest?
45	Patient	It started in my back and shot down my leg and ankle.	Patient	No.
46	Attending	Who is your family doctor?	Resident	Have you had any urine changes?
47	Patient	(The patient gives her doctor's name.)	Patient	No.
48	Attending	Have you ever seen a cardiologist?	Resident	Have you had any nausea or vomiting?
49	Patient	No.	Patient	Neither.
50		[Five lines removed for space.]	Resident	Have you had any constipation?
51			Patient	No.
52			Resident	Any fever or chill?
53			Patient	Maybe a fever. I've felt cold, too.

Emergency doctors in general are very experienced in addressing chest pain, and there is some degree of standardization (e.g., most doctors will ask about how the pain started, shortness of breath, and nausea). Despite this, there was considerable variation in the way that the doctors conducted their interviews. In the example above, the attending physician was very structured, progressing from questions about the patient's current symptoms (Lines 2-28), to past medical history (Lines 29-40), to finding out about contributors such as smoking (Lines 41-43), to finding out about other information such as who the patient's other doctors are (Lines 46-48). In contrast, the resident asked about current symptoms both at the beginning and the end of the interview (Lines 3-18 and Lines 40-52) and mixed questions about contributors such as travel with questions about the patient's medical history (Lines 20-39).

The attending asked fewer and more targeted questions overall than the resident did. The attending did ask general questions regarding the patient's chronic medical complaints (e.g., Lines 29 and 39) but asked specifically about problems potentially related to the current episode such as heart attack, stroke, or clots (Lines 33, 35, and 37). The resident did the opposite: asking about diabetes and high blood pressure specifically (Lines 24 and 26) but asking only a general question about cardiac issues (Line 20).

The different medical records systems used in each hospital also potentially affected the interview. The attending physician (in the hospital with paper records) had to ask about the patient's previous back pain diagnosis, using the patient as the link to past test results. The resident (in the hospital with electronic records) only needed to confirm that the patient was scheduled for a stress test the next day. However, the resident with access to more electronic documentation still asked more questions, implying that differences due to experience may outweigh the effects of the work system.

Attending physicians' interviews tended to be shorter and more structured than residents' interviews. In addition to these general differences, attending physicians tailored their interviews to respond to patient difficulty. The following example illustrates some of the differences between treating simple and more difficult patients (Table 5).

Table 5. Example of interviews for patients of different difficulty. The patients were both seen by the same attending physician (19 years of experience) at the suburban hospital during a shift with a difficulty score of 36. The difficult patient (left) complained of abdominal pain whereas the simpler patient (right) complained of shoulder pain.

	More Difficult Patient		Simple Patient	
Line	Speaker	Utterance	Speaker	Utterance
1	Attending	You have some belly pain, nausea, vomiting, and diarrhea?	Attending	Can you tell me about the pain?
2	Patient	It's actually in my back.	Patient	I thought I just slept on it wrong. I've been waking up numb and I have terrible shoulder pain. I can't feel my fingers at night.
3	Attending	Did it start there?		(The attending has the patient squeeze his fingers.)
4	Patient	Two days ago, I had a procedure with dye in my uterus.	Attending	Does that hurt your shoulder?
5	Attending	A hysteroscopy?	Patient	No.
6	Patient	That's it. They also took a biopsy of my cervix. A student did the procedure - I don't have a problem with that but I think that was part of it.		(The attending presses the patient's arm and shoulder.)
7	Patient	My cervix is upside down and that makes my uterus off so the cath couldn't go in. I don't have a tube because I've had four tubal pregnancies.	Patient	Pressing hurts it.
8	Patient	I wanted to see why my husband and I couldn't conceive. They found a block in the tube on my right side.		(The attending has the patient make his arm limp and moves it around.)
9	Patient	I get a sharp, shooting pain but it's not in my muscles.	Patient	It hurts less when I'm not using the muscles.
10	Attending	It's down deep, huh?		(The attending presses the patient's shoulder near his trapezius.)
11	Patient	Yeah, mostly on the left side. I took four ibuprofin this morning and laid down and I still felt bad. I got up a couple hours ago and lost bowel control. I could barely walk so I called the clinic and they said to go to the ED.	Attending	I'll x-ray the shoulder, but the rest looks fine.
12	Attending	Did the vomiting and diarrhea start today?		(The attending listens to the patient's chest.)
13	Patient	Yes.	Attending	Do you have any allergies?
14	Attending	Do you have any allergies?	Patient	Codeine.
15	Patient	Sulfa, and IV compazine/phenergan.	Attending	Have you been lifting anything?
16	Attending	Does it make you jumpy?	Patient	No.
17	Patient	Yes, but I can take phenergan orally.		
18	Attending	What about issues in your family?		
19	Patient	Blood pressure problems.		

The attending physician in this example obtained far more information from the more difficult patient than from the less difficult patient. The less difficult patient gave a brief description of their symptoms (Line 2) and the physician relied more on the exam to gather the information needed (Lines 3-12). Information such as the patient's vital signs and an EKG may have helped to eliminate more serious potential causes of shoulder pain, such as a heart issue. The more difficult patient gave much more detail about the background of their problem and symptoms (Lines 2-13). The patient also included more information about their past medical history.

Attending physicians also tended to shorten their interviews during busy shifts, whereas residents did not. This is not to say that residents did not adapt to workload; they may simply have adjusted in other ways, as demonstrated by the following quote from a resident:

> *(While looking at a new patient's chart) Time can impact how much I can look at a patient's chart before I go to see them, but I'm not busy now and I'm checking because the triage note said this has happened before.*

Goal Enactment

Quantitative Analyses of Goal Enactment

Residents. Residents had a mean goal enactment score of 0.14 ± 0.81. For residents alone, the ICC for the goal enactment behaviors was 0.26, indicating substantial between-doctor variance between residents on this factor. Unlike the resident-only analysis for the goal establishment factor, Level 1 analyses indicated that individual residents demonstrated more goal enactment behaviors with more difficult patients (Table 6). Level 2 analyses for intercept variance indicated that hospital predicted goal enactment behaviors such that residents at the urban hospital performed more goal enactment behaviors than residents at the suburban hospital.

Table 6. Multilevel Results for Residents on the Goal Enactment Factor

Model and Parameter	Parameter Estimate	*SE*	*t*	*df*	*p*
Level 1					
Patient difficulty	0.69	0.08	8.80	142	< 0.01
Level 2 (Intercept)					
Experience	0.14	0.18	0.78	14	0.45
Shift difficulty	-0.02	0.02	-0.99	14	0.34
Hospital	-0.51	0.18	-2.84	14	0.01
Level 2 (Slope)					
Experience	0.00	0.19	0.01	14	0.99
Shift difficulty	-0.02	0.02	-1.36	14	0.20
Hospital	-0.10	0.19	-0.55	14	0.59

Note. Predictors in **bold** are significant at $p < 0.05$.

Attendings. The attending physicians had a mean goal enactment score of -0.18 ± 0.59. For the attending physicians, the ICC for the goal enactment factor was 0.06, indicating no significant variance in the intercept on this factor and suspending further multilevel analysis. Because the ICC indicated that the nesting of patients within doctors was inconsequential, we used regression analysis to investigate the effects of our predictors on attending physicians' goal enactment behaviors. We entered our four predictors into a single model and found that individual attendings' goal enactment behaviors increased with increasing patient difficulty (Beta = 0.56, $p < 0.001$). We observed between-doctor differences such that attending physicians' goal enactment behaviors decreased with increasing shift difficulty (Beta = -0.20, $p = 0.004$). Hospital ($p = 0.19$) and experience ($p = 0.54$) did not predict goal enactment behaviors within the attending physicians. The overall model yielded an R^2 value of 0.36.

Qualitative Analysis of Goal Enactment

Although both residents and attending physicians adjusted goal enactment behaviors based on patient difficulty, only the residents adjusted their behaviors based on hospital. This is potentially due to difficulty in switching between multiple hospitals, as residents often had issues with hospital work practices. For example, one resident at the suburban hospital initially called the wrong doctor and had to change computers to resolve a prescription issue while caring for a patient. Similarly, a resident at the urban hospital initially went to the wrong room and later had to resolve an issue with an EKG for a patient.

In contrast to the residents, who sometimes struggled with specific work practices in a given hospital, attending physicians seemed to have developed more stable work behaviors. One attending physician stated that he used his own personal "workup template" for various complaints (chest pain, abdominal pain, etc.). He stated that he adds to the template for individual patients based on the physical exam, but he never subtracts from the template.

Uncertainty Reduction

Quantitative Analysis of Uncertainty Reduction

Uncertainty reduction was a relatively rare event (residents used uncertainty reduction behaviors for only 27 patients across the entire sample; attendings used uncertainty reduction behaviors for only 11 patients). Due to this, and the aforementioned issues with the distribution of the residuals, the following results should be interpreted with caution. Nevertheless, we include them here to guide further inquiry.

Residents. The mean for the residents' uncertainty reduction scores was 0.15 ± 1.09. The ICC for the uncertainty reduction factor for the residents was 0.10. None of the predictors used in this study were able to account for variance in the residents' uncertainty reduction behaviors at any level using multilevel modeling (all p values ≥ 0.35).

Attendings. Attending physicians had a mean uncertainty reduction score of -0.20 ± 0.53. The ICC for the uncertainty reduction factor for the attending physicians was 0.07. As with goal enactment, we therefore used regression to model attending physician behavior. We entered our four predictors into a single model and found that individual attendings' uncertainty reduction behaviors increased with increasing patient difficulty (Beta = 0.17, $p = 0.02$) and that more experienced attendings performed *more* uncertainty reduction behaviors (Beta = 0.15, $p = 0.048$). Hospital and shift difficulty did not predict uncertainty reduction behavior (both $p > 0.20$). However, the overall model yielded an R^2 value of only 0.06.

Qualitative Analysis of Uncertainty Reduction

Quantitative analyses indicated that residents used uncertainty reduction behaviors equally across several different contexts, but the attending physicians utilized uncertainty reduction behaviors only for more difficult cases. Pocket-sized medical references, internet searches, or data stored on smartphones serve as permanent information storage that the doctors can use to double-check themselves. One third-year resident even referred to her smartphone as her "external brain." For instance, one resident at the urban hospital with a high average uncertainty reduction score (1.21) looked up information regarding antibiotic dosages for several patients, none of whom were rated a three on patient difficulty. In contrast, an attending physician at the same hospital with a high average uncertainty reduction score (0.49) confronted with an overdose patient looked up information on the medication the patient took and how to reverse the effects but did not demonstrate uncertainty reduction for any of the other more straightforward patients.

Discussion

Emergency medicine requires physicians to manage simultaneously multiple patients of varying urgency within the constraints of different hospital-specific resources and the socio-culturally determined practice of Western medicine, including the recognition of patient demands. Our observational and analytical method allowed us to identify the influences of real-world contextual features on behavior in a quantifiable way, merging the strengths of naturalistic and quantitative approaches, while overcoming some of the limitations associated with each. Attending physicians and residents differed in goal establishment, goal enactment, and uncertainty reduction processes in response to both immediate (workload) and permanent (work system) differences in the work context. In light of these results, we revisit the primary issues raised in the introduction: the nature of expert knowledge, outcome versus process accounts of expertise, adaptation to context, and the development of expertise.

Nature of Knowledge

Differential sensitivity to patient difficulty, shift difficulty, and hospital among the attendings and residents reinforces our claim that the ability to recognize contextual features and adapt accordingly is critical to expertise (Shalin et al., 1997; Shalin & Bertram, 1996; Shalin & Verdile, 2003; Dubois & Shalin, 1995; Lippa & Shalin, 2016). Our data are potentially consistent with a template-based model of expert intuition and decision making (Gobet & Chassey, 2009; Tenison & Anderson, 2016), according to which patterns in the world give rise to networks (templates) composed of conceptual chunks which are themselves associated with various actions. Crucially, we do not conceptualize these as prescriptive knowledge structures, but rather distillations of the features that are relevant to the current situation.

Attending physicians' goal establishment behaviors were very sensitive to contextual features. We can envision an adaptation of templates with a core of essential information and many variable "slots"; Attendings may be better able to select which slots to fill and which to ignore when pressed for time. Attendings' adaptive ability may also reflect the more detailed mental models and enhanced ability to identify case typicality associated with expertise (Phillips et al., 2004), enabling more targeted adjustments to goal establishment behavior. Future research will need to address this possibility; goal establishment is an under-studied component of expertise that occurs in many domains including nursing, disaster response and military decision making. One contribution of the current effort is to operationalize and quantify real-world contextual features relevant to medicine (and potentially other domains) in order to better facilitate their incorporation in future research.

Outcome Versus Process Accounts of Expertise

Emergency medicine is an *open domain.* Outcomes are not necessarily attributable to the actions of a single physician. Further, desirable or realistic outcomes may differ across patients, precluding comparisons across doctors who encounter different cases in a shift. Moreover, the task is not simply to be "right" but to be efficient while managing limited resources and patient affect. Outcome measures are therefore difficult to quantify and evaluate across patients and doctors, given the adaptive goal establishment behavior we have documented. We thus favor the characterization of expertise as a process rather than an outcome. Our results indicate that experts and novices differ greatly in how they adapt their behavioral processes to changing contextual demands over immediate (patient difficulty), medium (shift difficulty) and long-term (hospital) time intervals. Outcome-based assessments such as measures of diagnostic or even treatment accuracy would not likely detect such differences, as these assessments are an incomplete reflection of care in the ED. This demands a change in the conceptualization of experimental tasks and measures purported to reveal expertise, as well as ways in which residents are trained and evaluated.

Despite our affinity for process measures, we have identified important gaps in the prevailing process accounts exemplified by Weiss and Shanteau (2014). Their valuable contributions do not specify the features that make cases similar, limiting our ability to compare processes across real-world cases. We have begun to identify features (such as patient characteristics, shift workload, and hospital work practices) that distinguish one situation from another, though it remains to be explored how experts identify the feature thresholds that trigger a change in behavior. Our findings improve the applicability of the process approach to naturalistic settings and facilitate the inclusion of such features into efforts to study and develop expertise.

Adaptation to Context

Context matters in the conceptualization of emergency medicine expertise. Consistent with Dreyfus and Dreyfus (2005), the ability to *modify* reasoning processes in response to local constraints appears to be a key feature of expertise. But this conclusion leaves researchers with the troublesome, albeit superficial inference that all behavior is unpredictable, and therefore psychological science is doomed (Brown et al., 1989; Suchman, 1993; Vera & Simon, 1993). We are not so gloomy, using our analysis to constrain the effects of context, and we include this in our process-oriented account of expertise and its acquisition.

One way in which we constrain the effects of context is to distinguish between two abstract frequency measures: goal establishment and goal enactment. Adaptive goal establishment is apparent in experts across all predictors. Adaptive goal enactment as measured by frequency is apparent in experts only for patient and shift difficulty, *but not hospital*. This is consistent with our previous emphasis on accepted socioculturally-constrained methods (Livingston, 1999) as the hallmark of disciplinary affiliation (Shalin et al., 1997). The medical system allows doctors to be relatively flexible when gathering information and determining a patient's problems; once a goal is identified, however, work practices and other constraints limit a doctor's flexibility (Shalin & Bertram, 1996). Different medications may be selected for a bacterial infection, for example, but a patient is likely to get some sort of pharmaceutical intervention. The small ICC values in the multilevel examination of attending goal enactment behaviors (in contrast to the residents) indicate that physicians' enactment behaviors become more similar to one another over time as they gain experience in the medical system. Qualitative evidence further supports this standardization account, as illustrated by the example of the attending who used a standardized workup template. Attending physicians may also have more established routines that stabilize their behavior compared to residents. The broader work system serves to reinforce consistency as well, evidenced by the availability of standardized sets of lab tests that doctors could order in both hospitals.

The second way in which we constrain the effects of context is to distinguish between the time constant of our predictors. We demonstrated the adaptive process of experienced attendings to both patient-level and shift-level workload—relatively short-term influences. Although researchers have described expert adaptation while solving a single problem over time (e.g., Ward et al., 2018), the adaptation of behavior across multiple similar situations described here is rarely discussed among the hallmark traits of expertise. Experts excel at the management of multiple patients—hardly a surprising conclusion, but completely absent from the contemporary conceptualization of medical expertise. The more streamlined interviews of the attending physicians likely facilitate adjustments in response to patient and shift difficulty compared to residents, who were perhaps still using algorithmic interview questions from their training rather than more personalized interviews developed with experience. Attending physicians also reserved uncertainty reduction for the more difficult patients, whereas residents did not show any systematic differences across patients. Attending physicians likely have less need to check relatively routine things like medication dosages.

In contrast, residents' response to shift and patient difficulty was more limited. They responded to patient difficulty but not shift difficulty, and only on goal enactment. Residents may claim to adjust to shift difficulty, as indicated by the quote from the resident who appeared to reduce chart reading if shift difficulty increased. However, these adaptations may be counterproductive if the lack of background information from the chart forces a longer patient interview.

Hospital represents a predictor with a longer-term time constant. The effect of hospital on physician behavior depends on expertise and the measure (goal establishment versus goal enactment). Attending physicians respond to hospital with adjustments in goal establishment, but not goal enactment. Residents respond to hospital with adjustments in goal enactment, but not goal establishment. Hospital is a proxy for both patient cohort and the resources available for patient care, including the availability of follow up care and the role of technology in patient management. The hospital effect on goal establishment provides a further example of how expert physicians are sensitive to differences in patient population and are able to take advantage of the flexibility allowed by a given work system. Attendings in the suburban hospital demonstrated more goal establishment behaviors than in the urban hospital. The T-sheets™ used at the suburban hospital to document the patient's care likely served as reminders of what to ask the patient or what the hospital expected to be documented; the T-sheets™ may predetermine many of the questions asked by the physicians. Doctors using electronic records at the urban hospital may have been less constrained, allowing them to ask only the questions they thought were necessary rather than fill out an entire form. These findings are consistent with the suggestion that the type of medical record system in use may affect the work practices of residents and attending physicians differently (Park, Lee, & Chen, 2012).

The presence of a hospital effect on goal enactment for the residents is interesting given the lack of such an effect for attending physicians, or for residents on the goal establishment factor. This may be a product of an inadequate stopping rule during goal establishment, leading residents to attempt to do too much in the urban hospital where more patients may need care that an emergency setting is not meant to provide. Nevertheless, the residents' goal enactment sensitivity to hospital suggests an adaptive process that the attending physician has largely accomplished. Residents make time-consuming, but otherwise inconsequential errors that increase enactment effort, such as knowing who to call, computer capabilities, and room layout.

Implications for the Development of Expertise

Contextual flexibility among the physicians in our sample appears to develop after residency is complete, beyond the boundaries of formal training. The development of context sensitivity evident in our sample of attending physicians appears to occur largely in the absence of systematically manipulated contexts and knowledge of outcomes, which even if known result from multiple interventions and stochastic processes. This obviates the deliberate practice paradigm with its reliance on carefully crafted training sessions and reinforces our focus on process. Relying on the traditional psychological method of expert-novice comparison, we consider the above assertions separately, including the evidence we provide and its theoretical significance regarding the conceptualization of expertise.

Expert skill development is often described as a progression from inflexible rule-based thinking to contextually-driven flexibility, facilitated in large part by deliberate practice (Dreyfus & Dreyfus, 2005; Ericsson, 2006). Emergency medical training in the United States is consistent with many elements of deliberate practice, including supervised practice of skills in medical school and progression towards independence under the guidance of more senior physicians during residency. However, deliberate practice is unlikely to account for the development of adaptive skill described in this study, particularly for goal establishment. Though residents are able to practice simulated care scenarios via training simulations and board exam preparation, such training rarely incorporates the variation in work practice and fluctuation in workload that affected physician behaviors in the present study. Further, real-world knowledge of patient outcomes is rare, and other objective metrics (e.g., length of patient interview) are not easily obtained in the clinical setting by the clinicians themselves.

Residents' goal establishment behaviors may be critiqued during simulation training or oral board practice but are rarely subject to feedback during real-world conditions in the same way as goal enactment behaviors. Residents work under the supervision of attending physicians who must concur with any treatments or interventions selected by the resident. However, attending physicians rarely accompany residents when interviewing patients. Attendings therefore give residents regular feedback on their goal enactment decisions but are rarely able to offer advice on goal establishment. Goal enactment behaviors also have a more tangible cost (in terms of money, resources, and patient outcome), warranting a focus in a clinical training setting. In addition, feedback occurs from other participants in the system such as the charge nurse (who may complain about bed utilization or availability), admitting physicians (who may resist taking a patient), or patients themselves who may return after discharge.

Goal establishment, on the other hand, may only warrant feedback in the case of excessively long interviews or incidentally and relatively rarely as part of a separate conversation (as in the cocaine example above). In addition, the residents in our sample rotated through several different hospitals (vs. attendings who typically only worked in one or two hospitals). Qualitative data indicated that residents' sensitivity to hospital on the goal enactment factor may have been due to this rotation; rather than adaptively changing goal enactment across hospital, the residents may simply have had difficulty navigating new work contexts so frequently. This rotation also potentially limited the influence of any specific work system on the patient interview by making it difficult or impractical to adjust before moving to the next hospital. Using a one-size-fits-all interview may be advantageous in such circumstances; physicians may be able to quickly adapt to the particulars of their final workplace when they finish training and start practicing independently.

Our findings suggest that training and assessment may need to be expanded to account for additional aspects of expert skill. Consistent with Bell et al. (2017), we believe more research is necessary to identify the relevant features of the work environment and how experts develop attunement to such features so that the acquisition of context sensitivity may be incorporated into the learning process. Doing so would allow physicians to learn to work efficiently in a variety of contexts *before* beginning to provide care independently. Although simulation-based training and oral board practice provide opportunities for feedback on goal establishment prior to independent practice, incorporating more opportunities for feedback about goal establishment during clinical duties could potentially speed acquisition of context sensitivity (as it may have done for patient-related adjustments to goal enactment behaviors). Exposure to real-world features in the scenarios encountered during simulations would also likely foster such sensitivity (Patterson, R., et al., 2010; Patterson, E., et al., 2010; Phillips et al., 2004). An additional contribution of this study is to begin characterizing relevant domain features to facilitate the design of such scenarios. In addition, assessments of skill such as those for licensure or to assess progress through training should incorporate a variety of situations to determine whether the student is able to adapt reasoning processes in response to different contextual features.

One hypothesis offered to explain the acquisition of skill in low-feedback environments is that unsupervised implicit learning may help develop tacit knowledge, and even potentially promote the acquisition of more complex explicit knowledge or skill (Patterson, R., et al., 2010). Immersive environments such as simulations may help promote such learning but assumed critical aspects for a given domain influence the design of simulation scenarios.

Strengths and Limitations

The primary strength of the study was the setting in which observations took place. The natural setting allowed the observer to record behaviors as influenced by real-world conditions, over multiple time scales. The second strength of the study was the comparison of attending physicians to residents. Residents are not true novices; they have medical degrees and those in our sample had at least one prior year of clinical experience. This study therefore provides insight into changes in decision making due to factors beyond the acquisition of didactic knowledge and basic technical skill. A unique advantage of the teaching environment was that residents naturally verbalized their thought processes when reporting to their attending physician.

However, using a natural setting introduces several limitations common to any observational study. We utilized a relatively small sample size for multilevel modeling, and our calculation for shift difficulty lacked temporal specificity (i.e., patients seen late in a shift would not affect patients discharged earlier in the shift, but both patients were associated with the same shift difficulty score). Further, our use of handwritten notes may have led to missed data during rapid conversational exchanges, though we have no reason to suspect any systematic differences across the groups or variables under study (for a more thorough discussion of these limitations, see Robinson et al., 2020). Many of these limitations are fundamental to the nature of our questions: Context sensitivity cannot be investigated without context, and in situ data collection is enormously time consuming.

The work practices described here are also potentially unique to the hospitals under observation and have since been modified. The T-Sheets™ used in one hospital have been replaced with an electronic system. Though our broader point about the effects caused by different work systems remains, some of the specific differences identified may no longer hold. It is even conceivable that a switch from paper to electronic records systems favors residents, who are more likely than attendings to have experienced such systems throughout their training.

Additional limitations are specific to the current findings. Attending physicians do not report to anyone. Behaviors related to goal enactment (e.g., evidence evaluation) may therefore have been more easily captured in the observations for the residents than the attending physicians. This will need to be explored further in subsequent analyses, perhaps by examining the differences between second- and third-year residents and fellows rather than between residents and attending physicians. Further, our data collection efforts were focused on the physicians' observed behavior rather than the actions and desires of the patients aside from noting whether the patient was cooperative. We have written about the contribution of the patient to medical decision making elsewhere (Lippa & Shalin, 2016; Lippa et al., 2017). However, our physician-centric perspective may not capture important patient considerations.

Another limitation of this study is the nature of our measures. The factor scores used as criterion measures excluded individual behaviors based on the results of the factor analysis. The absence of a specific behavior from a larger conceptually meaningful group does not mean that it was unimportant. We also reiterate that the individual behaviors were identified by a single rater who then recoded the observations with a delay to check for reliability.

We used years of experience to operationalize expertise despite known limitations (Ericsson, 2006; Rassafiani, 2009). However, we lacked the consistency of scenarios to apply methods endorsed by Weiss and Shanteau (2014), and we lacked access to patient records for each physician to classify expertise based on prior objective outcomes (Ericsson, 2007). The patterns of behaviors described above thus cannot be conclusively linked to varying degrees of medical skill. Nevertheless, we must first identify the characteristics of experts to operationalize expertise, and the experience-based comparisons used here permitted the identification of skills that appear to differentiate between doctors.

Finally, this study could not assess the mechanisms by which attending physicians acquired the observed contextual flexibility. Immersive residency apparently is not sufficient to allow residents to attain the same sensitivity as attending physicians. It remains undetermined what learning processes (implicit or otherwise) are most relevant for the development of such sensitivity, and how available feedback mechanisms are linked to the specific processes studied here. Continuing professional development in the form of literature review, informal consultation with other physicians, reflective practice, or other parallel activities may all contribute to the development of adaptive skill among attending physicians and are worthy of future examination. Further, although our analysis helps establish the contribution of contextual sensitivity to expert behavior, we do not address the cognitive processes by which this adaptation is achieved in real-time (i.e., deliberate or intuitive processes). Just as importantly, the adaptive behaviors demonstrated by the attending physicians on goal establishment processes appear to develop outside the dominant self-structured deliberate practice account of expert development. The study of expert skill acquisition would benefit from an expansion into more open domains to determine what skills can and cannot be accounted for by deliberate practice, and to identify additional contributing mechanisms such as semi-supervised conceptual pattern recognition processes. Future laboratory-based research will be required to explore the cognitive mechanisms underlying the observed adaptivity.

Conclusions

Rather than focus on the care *outcome*, we have examined the care *process* and documented that experts are able to adjust both goal establishment and goal enactment behaviors dynamically in response to situational factors. Indeed, we found that attending physicians are far more sensitive than residents to contextual features when establishing patient goals and are sensitive to different aspects of context when enacting those goals. Critically, the patients in our sample did not represent rare or especially difficult problems. Experts appear not only to adapt behavioral processes within single difficult cases (Ward et al., 2018), but also across multiple nominally similar, common cases. Differential sensitivity to context is an intriguing, understudied component of expert skill.

Acknowledgement

We are grateful to all of the physicians, patients, and staff who allowed us into their spaces to conduct our research.

Authors' Declarations

The authors declare that there are no personal or financial conflicts of interest regarding the research in this article.

The authors declare that they conducted the research reported in this article in accordance with the Ethical Principles of the Journal of Expertise.

This work utilizes a novel dataset combined with data analyzed in a prior manuscript (Robinson et al., 2020). The authors declare that they are not able to make the dataset publicly available but are able to provide it upon request.

ORCID iDs

F. Eric Robinson
https://orcid.org/0000-0002-5723-5658

Valerie L. Shalin
https://orcid.org/0000-0001-8135-2793

Debra Steele-Johnson
https://orcid.org/0000-0002-0314-1023

Brian Springer
https://orcid.org/0000-0002-7533-0464

References

Alqahtani, D., Rotgans, J., Mamede, S., Mahzari, M., Al-Ghamdi, G., & Schmidt, H. (2018). Factors underlying suboptimal diagnostic performance in physicians under time pressure. *Medical Education.* https://doi.org/10.1111/medu.13686.

Anderson, J. R. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Ash, J. S., Sittig, D. F., Campbell, E. M., Guappone, K. P., & Dykstra, R. H. (2007). Some unintended consequences of clinical decision support systems. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2007, 26–30.

Azer, S. A. (2007). Twelve tips for creating trigger images for problem-based learning cases. *Medical Teacher, 29*(2–3), 93–97.

Azer, S. A., Peterson. R., Guerrero, A.P., & Edgren, G. (2012). Twelve tips for constructing problem-based learning cases. *Medical Teacher, 34*(5), 361–367.

Bell, B. S., Tannenbaum, S. I., Ford, J. K., Noe, R. A., & Kraiger, K. (2017). 100 years of training and development research: What we know and where we should go. *Journal of Applied Psychology, 102*(3), 305-323. https://doi.org/10.1037/apl0000142

Benner, P. (2004). Using the Dreyfus model of skill acquisition to describe and interpret skill acquisition and clinical judgment in nursing practice and education. *Bulletin of Science, Technology & Society, 24*(3), 188-199.

Bliese, P. D. (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.

Brooks, L., Norman, G., & Allen, S. (1991). Role of specific similarity in a medical diagnostic task. *Journal of Experimental Psychology: General, 120*(3), 278-287. https://doi.org/10.1037//0096-3445.120.3.278

Brown, J., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42. https://doi.org/10.3102/0013189X018001032

Carbonell, K., Stalmeijer, R., Konings, K., Segers, M., & van Merrienboer, J. (2014). How experts deal with novel situations: A review of adaptive expertise. *Educational Research Review, 12,* 14-29. https://doi.org/10.1016/j.edurev.2014.03.001

Catchpole, K., & Alfred, M. (2018). Industrial conceptualization of health care versus the naturalistic decision making paradigm: Work as imagined versus work as done. *Journal of Cognitive Engineering and Decision Making, 12*(3), 222-226. https://doi.org/10.177/1555343418774661

Cheng, P.W. (1985). Restructuring versus automaticity: Alternative accounts of skill acquisition. *Psychological Review, 92*(3), 414-423. https://doi.org/10.1037/0033-295X.92.3.414

Cognition and Technology Group at Vanderbilt (1993). Designing learning environments that support thinking: The Jasper Series as a case study. In T. M. Duffy, J. Lowyck, & D. H. Jonassen (Eds), *Designing environments for constructive learning* (pp. 9-36). Berlin: Springer-Verlag.

Croker, A., Loftus, S., & Higgs, J. (2008). Multidisciplinary clinical decision making. In J. Higgs, M. Jones, S. Loftus, & N. Christensen (Eds.) *Clinical reasoning in the health professions* (pp. 291-298). Elsevier: Boston.

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

Creswell, J. (2007). *Qualitative inquiry and research design: Choosing among five approaches, 2nd edition.* Sage Publications, Inc.

Delaney, P. (2018). The role of Long-Term Working Memory and Template Theory in contemporary expertise research. *The Journal of Expertise, 1*(3), 155-161.

Donner-Banzhoff, N., Seidel, J., Sikeler, A., Bösner, S., Vogelmeier, M., Westram, A., Feufel, M., Gaissmaier, W., Wegwarth, O., & Gigerenzer, G., 2017). The phenomenology of the diagnostic process: A primary care-based survey. *Medical Decision Making, 37*(1), 27-34. https://doi.org/10.1177/0272989X16653401.

Dreyfus, H. L., & Dreyfus, S. E. (2005). Expertise in real world contexts. *Organization Studies, 26*(5), 779-792. https://doi.org/10.1177/0170840605053102

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* (pp. 167-250). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (eds), *The Cambridge handbook of expertise and expert performance* (pp. 683-703). New York, NY: Cambridge University Press.

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: The study of clinical performance. *Medical Education, 41,* 1124-1130. https://doi.org/10.1111/j.1365-2923.2007.02946.x

Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: A general overview. *Academic Emergency Medicine, 15*, 988-994. https://doi.org/10.1111/j.1553-2712.2008.00227.x

Ericsson, K. A. (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*(4), 458-463. https://doi.org/10.1002/acp.3029

Falzer, P. (2018). Naturalistic decision making and the practice of health care. *Journal of Cognitive Engineering and Decision Making*, 12(3), 178-193. https://doi.org/10.1177/1555343418773915

Forsythe, D. E. (1993). The construction of work in artificial intelligence. *Science, Technology, & Human Values, 18*(4), 460-479.

Foy, R., Hempel, S., Rubenstein, L., Suttorp, M., Seelig, M., Shanman, R., & Shekelle, P. (2010). Meta analysis: Effect of interactive communication between collaborating primary care physicians and specialists. *Annals of Internal Medicine, 152*(4), 247-258. https://doi.org/10.7326/0003-4819-152-4-201002160-00010.

Franklin, A., Robinson, D., & Zhang, J. (2014). Characterizing the nature of work and forces for decision making in emergency care. In V. Patel, D. Kaufman, & T. Cohen (Eds.), *Cognitive informatics in health and biomedicine* (pp. 1-13). Springer-Verlag.

Gobet, F., & Chassy, P. (2009). Expertise and intuition: A tale of three theories. *Minds and Machines, 19,* 151-180. https://doi.org/10.1007/s11023-008-9131-5

Gobet, F., Lloyd-Kelly, M., & Lane, P. (2017). Computational models of expertise. In D. Hambrick, G. Campitelli, & B. Macnamara (Eds.), *The science of expertise* (pp. 347-364). New York: Psychology Press.

Gobet, F., & Simon, H. (1996). Templates in chess memory: A mechanism for recalling several boards. *Cognitive Psychology, 31*(1), 1-40. https://doi.org/10.1006/cogp.1996.0011.

Gore, J., Banks, A., & McDowall, A. (2018). Developing cognitive task analysis and the importance of socio-cognitive competence/insight for professional practice. *Cognition, Technology, & Work, 20,* 555-563. https://doi.org/10.1007/s10111-018-0502-2

Gorsuch, R. (1974). Factor analysis. W.B. Saunders Company.

Hoffman, R. R., Coffey, J. W., Ford, K. M. and Novak, J. D. (2006). A method for eliciting, preserving, and sharing the knowledge of forecasters. *Weather and Forecasting, 21*, 416-428.

Hutchins, E. (2005). Material anchors for conceptual blends. *Journal of Pragmatics, 37*(10), 1555-1577. https://doi.org/10.1016/j.pragma.2004.06.008

Ibrahim M. E., Al-Shahrani, A. M., Abdalla, M. E., Abubaker, I. M., & Mohamed, M. E. (2018). The effectiveness of problem-based learning in acquisition of knowledge, soft skills during basic and preclinical sciences: Medical students' points of view. *Acta Informatica Medica*, *26*(2), 119–24.

Johnson, J., Miller, S., & Horowitz, S. (2008). Systems-based practice: Improving the safety and quality of patient care by recognizing and improving the systems in which we work. In K. Henriksen, J. Battles, M. Keyes, & M. Grady (Eds.) *Advances in patient safety: New directions and alternative approaches (Vol. 2: Culture and redesign).* Rockville, MD: Agency for Healthcare Research and Quality.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515-526.

Klein, G. A. (1989). Recognition-primed decisions. In W. B. Rouse (Ed.), *Advances in man-machine systems research* (Vol. 5, pp. 47–92). Greenwich, CT: JAI Press.

Klein, G. (2007a). Flexecution as a paradigm for replanning, part 1. *IEEE Human Centered Computing, 22*(5), 79-83. https://doi.org/10.1109/MIS.2007.4338498

Klein, G. (2007b). Flexecution as a paradigm for replanning, part 2. *IEEE Human Centered Computing*, *22*(6), 108-112. https://doi.org/10.1109/MIS.2007.107

Klein, G. (2015). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition, 4*, 164-168.

Knief, U., & Forstmeier, W. (2020). Violating the normality assumption may be the lesser of two evils. *bioRxiv*. https://doi.org/10.1101/498931.

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods*, *12*(3), 418-435. https://doi.org/10.1177/1094428107308984

Laird, J., Newell, A., & Rosenbloom, P. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence, 33*(1), 1-64. https://doi.org/10.1016/0004-3702(87)90050-6

Lave, L. & Wenger, E. (1991). Situated learning: Legitimate peripheral participation. Cambridge: Cambridge University Press.

Lippa, K., Feufel, M., Robinson, F., & Shalin, V. (2017). Navigating the decision space; Shared medical decision making as distributed cognition. *Qualitative Health Research, 27*(7), 1035-1048.

Lippa, K.D . & Shalin, V. L. (2016). Creating a common trajectory: Shared decision making and distributed cognition in medical consultations. *Patient Experience*, *3*(2), 73-82. https://doi.org/10.35680/2372-0247.1116

Livingston, R. (1999). Cultural issues in diagnosis and treatment of ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry, 38*(12), 1591–1594. https://doi.org/10.1097/00004583-199912000-00021

LeCun, Y. (2017, September 6-8). *How does the brain lean so much so quickly?* [Conference presentation]. Cognitive Computational Neuroscience, New York, NY, United States. https://www.youtube.com/watch?v=cWzi38-vDbE

Martin, L., Haskard-Zolnierek, K., & DiMatteo, M. R. (2010). Health behavior change and treatment adherence. New York, NY: Oxford University Press.

MERIT Study Investigators (2005). Introduction of the medical emergency team (MET) system: A cluster-randomized controlled trial. *The Lancet, 365*(9477), 2091-2097

Mogford, R., Allendoerfer, K., Snyder, M., Hutton, R., & Rodgers, M. (1997). Application of the recognition-primed decision model to the study of air traffic controller decision making. Proceedings of the Ninth International Symposium on Aviation Psychology, 1, 739-744.

Moran-Barrios, J., & Gauna-Bahillo, P. (2010). Reinventing specialty training of physicians? Principles and challenges. *Nefrologia, 30*(6), 604-612. https://doi.org/10.3265/Nefrologia.pre2010.Jul.10559

Myneni, S., Cohen, T., Almoosa, K., & Patel, V. (2014). Standard solutions for complex settings: The idiosyncrasies of a weaning protocol used in practice. In V. Patel, D. Kaufman, & T. Cohen (Eds.), *Cognitive informatics in health and biomedicine* (pp. 183-202). Springer-Verlag.

Norman, D. A. (1986) User-centered system design: New perspectives on human-computer Interaction. In: Norman, D.A. and Draper, S.W., Eds., *Cognitive engineering*, Lawrence Erlbaum Associates, Hillsdale, 31-61.

Park, S., Lee, S., & Chen, Y. (2012). The effects of EMR deployment on doctors' work practices: A qualitative study in the emergency department of a teaching hospital. *International Journal of Medical Informatics, 81*(3), 204-217. https://doi.org/10.1016/j.ijmedinf.2011.12.001

Patel, V., Cohen, T., Batwara, S., & Almoosa, K. (2014). Teamwork and error management in critical care. In V. Patel, D. Kaufman, & T. Cohen (Eds.), *Cognitive informatics in health and biomedicine* (pp. 1-13). Springer-Verlag.

Patel, V., & Groen, G. (1986). Knowledge based solution strategies in medical reasoning. *Cognitive Science, 10*(1), 91-116. https://doi.org/10.1207/s15516709cog1001_4

Patel, V., Groen, C., & Patel, Y. (1997). Cognitive aspects of clinical performance during patient workup: The role of medical expertise. *Advances in Health Sciences Education, 2*(2), 95-114. https://doi.org/10.1023/A:1009788531273

Patel, V., Kaufman, D., & Cohen, T. (2014). Complexity and errors in critical care. In V. Patel, D. Kaufman, & T. Cohen (Eds.), *Cognitive informatics in health and biomedicine* (pp. 1-13). Springer-Verlag.

Patel, V., Kaufman, D., & Kannampallil, T. (2013). Diagnostic reasoning and decision making in the context of health information technology. *Reviews of Human Factors and Ergonomics, 8*(1), 149-190. https://doi.org/10.1177/1557234X13492978

Patterson, R., Pierce, B., Bell, H., & Klein, G. (2010). Implicit learning, tacit knowledge, expertise development, and naturalistic decision making. *Journal of Cognitive Engineering and Decision Making*, 4(4), 289-303. https://doi.org/10.1518/155534310X12895260748867

Patterson, E., Roth, E., & Woods, D. (2010). Facets of complexity in situated work. In E. Patterson & J. Miller (Eds.), Macrocognition metrics and scenarios: Design and evaluation for real-world teams. Ashgate.

Payne, V., & Patel, V. (2014). Enhancing medical decision making when caring for the critically ill: The role of cognitive heuristics and biases. In V. Patel, D. Kaufman, & T. Cohen (Eds.), *Cognitive informatics in health and biomedicine* (pp. 203-231). Springer-Verlag.

Phillips, J., Klein, G., & Sieck, W. (2004). Expertise in judgment and decision making: A case for training intuitive decision skills. In D. Koehler & N. Harvey (Eds.), Blackwell handbook of judgment and decision making. Blackwell Publishing Ltd.

Preeti B, Ashish A, & Shriram G. (2013). Problem based learning (PBL): An effective approach to improve learning outcomes in medical teaching. *Journal of Clinical and Diagnostic Research*, 7(12), 2896–7.

Putnam, H. (1979). Philosophical papers: Volume 2, mind, language and reality. Cambridge: Cambridge University Press.

Quine, W. (2008). Chapter 31: Three indeterminacies. In D. Follesdal & D. Quine (Eds.), *Confessions of a confirmed extensionalist and other essays* (pp. 368-386). Harvard University Press.

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

Rassafiani, M. (2009). Is length of experience an appropriate criterion to identify level of expertise? *Scandinavian Journal of Occupational Therapy, 16*(4), 247-256. https://doi.org/10.3109/11038120902795441

Reason, J. (1990). Human error. Cambridge University Press.

Reitman, W. R. (1964). Heuristic decision procedures, open constraints, and the structure of ill-defined problems. In M. W. Shelly & G. L. Bryan (Eds.), *Human judgements and optimality*. N.Y.: John Wiley and Sons.

Robinson, F. E., Feufel, M., Shalin, V., Steele-Johnson, D., & Springer, B. (2020). Rational adaptation: Contextual effects in medical decision making. *Journal of Cognitive Engineering and Decision Making, 14*(2), 112-131. https://doi.org/10.1177/1555343420903212.

Roe, R., Busemeyer, J., & Townsend, J. (2001). Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychological Review, 108*(2), 370-392.

Roijers, D. M., Vamplew, P., Whiteson, S., & Dazeley, R. (2013). A survey of multi-objective sequential decision-making. J*ournal of Artificial Intelligence Research, 48*, 67-113.

Schmidt, H., van Gog, T., Klein Nagelvoort-Schuit, S., van den Berge, K., van Daele, P., Bueving, H., van der Zee, T., van den Broek, W. W., van Saase, J. L., & Mamded, S. (2017). Do patients' disruptive behaviours influence the accuracy of a doctor's diagnosis? A randomised experiment. *BMJ Quality and Safety, 26*(1), 19-23. https://doi.org/10.1136/bmjqs-2015-004109.

Schraagen, J. M., Chipman, S. F., & Shalin, V. L. (Eds.). (2000). Cognitive task analysis. Lawrence Erlbaum Associates Publishers.

Servant-Miklos, V. F. C (2018). Problem solving skills versus knowledge acquisition: The historical dispute that split problem-based learning into two camps. *Advances in Health Sciences Education Theory and Practice*. https://doi.org/10.1007/s10459-018-9835-0.

Shalin, V., & Bertram, D. (1996). Functions of expertise in a medical intensive care unit. *Journal of Experimental and Theoretical Artificial Intelligence, 8*, 209-227. https://doi.org/10.1080/095281396147302

Shalin, V. L. , Geddes, N. D., Bertram, D. L., Szczepkowski, M. & DuBois, D. (1997). Expertise in dynamic, physical task domains. In P. Feltovich, K. Ford, & R. Hoffman (eds). *Expertise in context: Human and machine*. Menlo Park, CA: AAAI Press.

Shalin, V. L. & Verdile, C. L. (2003). The identification of knowledge content and function in manual labor. *Ergonomics, 46*(7), 695-713. https://doi.org/10.1080/0014013031000085626

Suchman, L. (1993). Response to Vera & Simon's situated action: A symbolic interpretation. *Cognitive Science, 17*(1), 71-75. https://doi.org/10.1016/S0364-0213(05)80011-4

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. *Cognition and Instruction, 12*(3), 185-233. https://doi.org/10.1207/s1532690xci1203_1

Tenison, C., & Anderson, J. (2016). Modeling the distinct phases of skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42*(5), 749-767. https://doi.org/10.1037/xlm0000204

Vallacher, R., & Wegner, D. (2012). Action identification theory. In P. Van Lenge, A. Kruglanski, & E.T. Higgins (eds.), *Handbook of theories of social psychology* (Vol. 1). London: Sage Publishing.

Van Lehn, K. (1987). Learning one subprocedure per lesson. *Artificial Intelligence, 31*(1), 1-40. https://doi.org/10.1016/0004-3702(87)90080-4

Vankipuram, M., Ghaemmaghami, V., & Patel, V. (2014). Adaptive behaviors in complex clinical environments. In V. Patel, D. Kaufman, & T. Cohen (Eds.), *Cognitive informatics in health and biomedicine* (pp. 147-182). Springer-Verlag.

Vera, A. H., & Simon, H. A. (1993). Situated action: A symbolic interpretation. *Cognitive Science, 17*(1), 7–48. https://doi.org/10.1207/s15516709cog1701_2

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

Vong, W., Navarro, D., & Perfors, A. (2016). The helpfulness of category labels in semi-supervised learning depends on category structure. *Psychonomic Bulletin & Review, 23*, 230-238. https://doi.org/10.3758/s13423-015-0857-9

Voss, J. F., Greene, T. R., Post, T. A., & Penner, B. C. (1983). Problem-solving skill in the social sciences. In Psychology of learning and motivation (Vol. 17, pp. 165-213). Academic Press.Ward, P., Gore, J., Hutton, R., Conway, G., & Hoffman, R. (2018). Adaptive skill as the *conditio sine qua non* of expertise. *Jounal of Applied Research in Memory and Cognition, 7*(1), 35-50. https://doi.org/10.1016/j.jarmac.2018.01.009

Weiss, D., & Shanteau, J. (2014). Who's the best? A relativistic view of expertise. *Applied Cognitive Psychology*, *28*(4), 447-457. https://doi.org/10.1002/acp.3015

Wong, B. L. W., & Kodagoda, N. (2015). How analysts think: Inference making strategies. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59(1), 269–273. https://doi.org/10.1177/1541931215591055

Wong, B. L. W., Sallis, P., & O'Hare, D. (1995). Information portrayal for decision support in dynamic intentional process environments (Information Science Discussion Papers Series No. 95/6). University of Otago. Retrieved from http://hdl.handle.net/10523/955

Zhang, J., & Norman, D. (1995). A representational analysis of numeration systems. *Cognition*, *57*(3), 271-295. https://doi.org/10.1016/0010-0277(95)00674-3

Received: 6 February 2022
Revision received: 8 June 2022
Accepted: 28 July 2022

Appendix

Coding Manual Used for Qualitative Data Coding

Behavior types and subtypes:

1. *Information gathering behavior.* These behaviors were broken down into both the source of the information and the type of information. A given piece of information was associated with both a source and type. For example, if the patient told the doctor that they had a headache, the source of information would be coded as the patient and the type of information would be current symptoms.
    a. Source
        i. Exams
        ii. Tests/images
        iii. The patient.
        iv. Patient's family/friends
        v. Medical records
        vi. Hospital staff
        vii. Internet/references
        viii. Miscellaneous.
    b. Type
        i. Current symptoms (directly related to the course of the present illness, describes the patient's experience of their symptoms)
        ii. Timeline (how the current problem is progressing, how long the patient has had problems, etc.)
        iii. Past medical information (related to previous or ongoing/chronic medical issues)
        iv. Contributors (non-symptoms or non-medical factors that may be related to the current problem or may help explain it. Also includes things that may exacerbate conditions - risk factors, family history, diet, whether meds have been taken, context of injury, etc.)
        v. Reference (checking how drugs interact, finding proper dosages, getting background information about a disease or disorder, etc.)
        vi. Other (relevant information but not directly tied to the patient's illness – e.g., who the patient's family doctor is, etc.)
2. *Diagnostics.* These behaviors were intended to try to establish a specific cause for symptoms or to eliminate alternative causes. Examples of diagnostic behavior include trying treatments as diagnostic tools, eliminating alternative diagnoses or treatments, comparing hypotheses, stating a diagnosis, etc.
3. *Evidence evaluation behavior.* These behaviors assessed the information that had been gathered for quality and completeness. This category included critically examining assumptions or reliability of tests, limitations of tools, or reliability of other evidence like a patient's oral history. It also included debating how likely a hypothesis was or how well a hypothesis was supported by the data or if there was converging evidence. Considering mitigating factors such as the patient's baseline values also fell under this category. These behaviors were generally focused on determining what symptoms and lab values should be taken seriously or how seriously to take a finding.
4. *Patient management behavior.* These behaviors were intended to manage the patient and ensure their care progressed as smoothly as possible. Patient management behaviors were broken into four different types:

a. Collaboration (telling a patient what's going on, asking for questions, offering reassurances, getting things to make the patient comfortable, etc.)
b. Treatment (related to regulating a patient's condition, treating a patient, etc.)
c. Consulting (talking with others about how best to care for the patient)
d. Logistics (trying to maximize the patient's benefit within the medical system (such as insurance benefits), taking action to minimize patient stay or resources used, making sure a treatment is appropriate, etc.)

5. *System management behavior.* These behaviors were about adjusting to accommodate structural constraints. They included working within the context of the hospital structure and current conditions such as a busy lab, working within the constraints of the medico-legal structure, etc. Also included using workarounds or "tricks of the trade."
6. *Filtering behavior.* Filtering behavior helped to establish the scope of the problem the doctor had to solve. These behaviors limited the information evaluated by trying to explain findings and remove them from the current problem or tried to determine if there was another problem that needed to be addressed. For example, if a patient complaining of chest pain and a headache said they took nitro the doctor may explain the headache as a side effect of the nitro and not part of the chest pain.

<u>Guidelines:</u>

1. Everything contained in a given segment was coded. If something also occurred in a prior segment, it got counted in both segments. For exceptions, see the rest of the notes below.
2. We only coded actions performed by the person being shadowed. For instance, if a resident was shadowed and the attending displayed a "codable" behavior or was given information that the resident didn't know about we did not code that – we only coded behavior performed by or information given to the resident.
3. We counted each discrete piece of information separately. For example, if the triage note said "patient is allergic to Ativan and has acute panic disorder," it counted as 2 discrete pieces of information. A good rule of thumb was whether something could be broken into multiple sentences (e.g., The patient is allergic to Ativan. The patient has acute panic disorder.).
4. We wanted to avoid double counting the exact same thing. To avoid that, when dealing with question-and-answer interactions, we only coded the answer (the information that was transmitted). If the patient did not answer, we coded based on the question that was asked to help capture thought processes.
5. If the doctor asked questions such as asking the patient to name the current date, we counted that as an exam.
6. When the patient described generalities of an ongoing condition such as what their seizures are normally like, we coded it as past medical information. If they discussed what happened during that day's episode, we coded it as current symptoms.
7. Information about the timing of past procedures (e.g., when a patient had surgery) counted as past medical information (not timeline).
8. Information from EMTs was counted as coming from staff, even though EMTs don't technically work for the hospital. Information from the patient's regular doctor was counted as coming from a miscellaneous source.
9. Information checked *purely* for documentation purposes was not counted. For example, if the doctor double checked a lab value for documentation, this was not counted as using a lab test. The act of documentation itself was coded elsewhere.
10. Discharging or admitting the patient counted as treatment.