9-3-2014

Metacognition and Decision-Making Style in Clinical Narratives

Limor Hochberg

Department of Psychology, College of Liberal Arts

Rochester Institute of Technology

Metacognition and Decision-Making Style in Clinical Narratives

by

Limor Hochberg

A Thesis in

Applied Experimental & Engineering Psychology

Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

September 3, 2014

We approve the thesis of Limor Hochberg:

Dr. Esa Rantanen Date

Associate Professor, Dept. of Psychology, RIT

Faculty Advisor and Chair of the Thesis Committee

Dr. Cecilia O. Alm Date

Assistant Professor, Dept. of English, RIT

Co-Advisor and Reader

Dr. Anne Haake Date

Professor, Golisano College, RIT

Reader

Dr. Caroline M. DeLong Date

Associate Professor, Dept. of Psychology, RIT

Reader

Acknowledgments

I would like to thank my thesis faculty advisor and committee chair, Dr. Esa Rantanen, co-advisor, Dr. Cecilia O. Alm, and thesis committe members, Dr. Anne Haake and Dr. Caroline M. DeLong, for their continued support of my work. Thanks for your patience, advice, careful review, and questions along the way. I'd also like to thank the Human-Centered Computing Group, who welcomed me as a member and allowed me to work on previously generated data. I'd like to specifically thank Dr. Qi Yu, a member of the HCC group, who consulted on the computational modeling work.

I'd like to thank RIT for my graduate education in general, and the Psychology Department faculty and staff in particular. Thanks also to the following RIT institutions for providing funding towards software purchases and conference travel: the Department of Psychology, College of Liberal Arts, Office of Graduate Studies, and Division of Student Affairs. This research has also been supported by an RIT College of Liberal Arts Faculty Development Grant, a Xerox award, and NIH award R21 LM01002901.[1]

Parts of this work have been published in the proceedings of the 8th Linguistic Annotation Workshop (Hochberg et al., 2014a) and the BioNLP Workshop (Hochberg et al., 2014b).

Thanks to my wonderful annotators, Amir Sivan and Daniel Nystrom. Thanks for your hard work, patience and insight into physician decision-making.

Finally, thanks to my family, friends, and my wonderful husband. I couldn't have done it without your love, advice, and support.

[1] This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Abstract

Clinical decision-making has high-stakes outcomes for both physicians and patients, yet little research has attempted to model and automatically annotate such decision-making. The dual process model (Evans, 2008) posits two types of decision-making, which may be ordered on a continuum from *intuitive* to *analytical* (Hammond, 1981). Training clinicians to recognize decision-making style and select the most appropriate mode of reasoning for a particular context may help reduce diagnostic error (Norman, 2009).

This study makes preliminary steps towards detection of decision style, based on an annotated dataset of image-based clinical reasoning in which speech data were collected from physicians as they inspected images of dermatological cases and moved towards diagnosis (Hochberg et al., 2014a). A classifier was developed based on lexical, speech, disfluency, physician demographic, cognitive, and diagnostic difficulty features to categorize diagnostic narratives as intuitive vs. analytical; the model improved on the baseline by over 30%. The introduced computational model provides construct validity for the dual process theory. Eventually, such modeling may be incorporated into instructional systems that teach clinicians to become more effective decision makers.

In addition, metacognition, or self-assessment and self-management of cognitive processes, has been shown beneficial to decision-making (Batha & Carroll, 2007; Ewell-Kumar, 1999). This study measured physicians' metacognitive awareness, an online component of metacognition, based on the confidence-accuracy relationship, and also exploited the corpus annotation of decision style to derive decision metrics. These metrics were used to examine the relationships between decision style, metacognitive awareness, expertise, case difficulty, and diagnostic accuracy. Based on statistical analyses, intuitive reasoning was associated with greater diagnostic accuracy, with an advantage for expert physicians. Case difficulty was associated with greater user of analytical decision-making, while metacognitive awareness was linked to decreased diagnostic accuracy. These results offer a springboard for further research on the interactions between decision style, metacognitive awareness, physician and case characteristics, and diagnostic accuracy.

Contents

Acknowledgments **ii**

Abstract **iii**

Introduction **1**

Literature Review 2

What is metacognition? 2

Research on metacognition 4

Measurement of metacognition 5

Confidence judgments as measures of metacognitive awareness 6

Methodology for measurement of metacognitive awareness 8

Studies of verbal data in clinical reasoning 8

Decision-making style 10

The Cognitive Continuum framework 11

Purpose of the Research 13

Hypotheses 14

Contributions 15

Method **17**

Participants 17

Apparatus and Stimuli 17

Procedure 18

Preliminary Data Analysis 19

Corpus Annotation of Decision Style 20

Results **23**

Decision Style Annotation 23

Quantitative annotation analysis 23

Annotator strategy analysis 25
Narrative case study 26
Automatic Annotation of Decision Style 28
Data selection and labeling for computational modeling 28
Methods 30
Decision style modeling results 33
Physician-Level Metrics of Decision Style and Metacognitive Awareness 34
Physician profiles of decision style 34
Physician profiles of metacognitive awareness 35
Hypothesis Evaluation 36
Decision style 37
Metacognitive awareness 43
Decision style and metacognitive awareness 45
Discussion **48**
Decision Style Annotation and Modeling 48
Decision Style and Metacognitive Awareness 50
Decision style and diagnostic accuracy 50
Intuitive reasoning, diagnostic accuracy, and expertise 51
Case difficulty and decision style 52
Metacognitive awareness and diagnostic accuracy 52
Metacognitive awareness, diagnostic accuracy, and expertise 54
Metacognitive awareness and decision style 54
Limitations 55
Applications 57
Conclusion **59**
References **60**

Appendices **A1**

Appendix A A1

Appendix B A2

Appendix C A5

Appendix D A6

Appendix E A8

Appendix F A10

List of Figures

1 Schematic representation of the confidence-accuracy relationship 6

2 Sample image case 18

3 Overview of annotation methodology 20

4 The decision-making continuum and rating scale 21

5 Distribution of decision ratings on a 4-point scale 24

6 Distribution of decision ratings on a 2-point scale 24

7 Annotator reliability for the 4-point rating scale 24

8 Annotator reliability for the 2-point rating scale 24

9 Annotator reliability, as measured by linear weighted kappa 25

10 Distribution of extreme agreement ratings on duplicated narratives 27

11 Narrative case study 27

12 Narrative labeling pipeline 29

13 Initial corpus labeling 29

14 Center-of-gravity strategy 30

15 Distribution of physician decision scores by expertise 35

16 Distribution of metacognitive awareness 36

17 Physician correctness by decision style 40

18 Relationship between physician decision style and diagnostic accuracy . . . 41

19 Success in intuitive reasoning by expertise 42

20 Relationship between case decision score and case difficulty 43

21 Relationship between case difficulty and analytical reasoning 44

22 Relationship between metacognitive awareness and diagnostic accuracy . . . 45

23 Metacognitive awareness by expertise 45

24 Relationship between metacognitive awareness and physician decision style 46

25 Relationship between metacognitive awareness and analytical reasoning . . 47

A1 Concept map of major study variables and hypotheses A1

C1 Decision style annotation confusion matrix A5

F1 Distribution of physician decision scores by gender A10
F2 Diagnostic accuracy by gender A11
F3 Success in intuitive reasoning by gender A11
F4 Metacognitive awareness by gender A12

List of Tables

1 Distribution of Annotator Ratings 23

2 Annotator Use of Factors in Rating Decision Style 26

3 Class Label Statistics 31

4 Feature Types Used For Decision Style Modeling 32

5 Decision Style Classifier Performance on Final Test Set 33

6 Feature Combination Study: Performance on Final Test Set 34

7 Summary of Case and Physician Metrics 38

8 Distribution of Diagnostic Correctness by Decision Style 39

9 Schematic Representation of Decision Style by Correctness 39

D1 Annotator Strategy Questionnaire A6

D2 Annotator Strategy Questionnaire - Continued A7

E1 Complete List of Lexical Features Used For Decision Style Modeling A8

E2 Complete List of Lexical Features Used For Decision Style Modeling - Continued A9

Introduction

Clinical decision-making is a rich and complex cognitive process, with high stakes for both physicians and patients. Effective clinical decision-making is likely to be found in individuals who can accurately assess their knowledge base and monitor, evaluate, and implement changes to cognitive strategies. This general construct of self-assessment is known as *metacognition*, and it is one of the key variables of interest in the current study. This is especially relevant for environments characterized by high validity (e.g., statistically regular environments, from which predictive cues may be learned), such as medicine (Kahneman & Klein, 2009).

Another factor of interest is clinical decision-making style. What *decision style* is a physician using in a given situation? Is the decision based more on gut instinct and immediate recognition, or careful lists of possibilities and subsequent logical analysis? The former is known as intuitive decision-making, and the latter as analytical decision-making. This distinction is drawn by Kahneman's heuristics and biases framework (Tversky & Kahneman, 1974), as well as dual-process models of reasoning (Evans, 2003; Kahneman & Frederick, 2002; Stanovich & West, 2000). Thus *decision style*, as viewed through the lens of this framework, is the other key variable of interest in the current investigation.

Metacognition has been shown to aid decision-making (Batha & Carroll, 2007; Ewell-Kumar, 1999). In addition, metacognitive processes, such as those responsible for detecting pattern anomalies, regulate the switch from intuitive to analytical modes of cognition during the decision-making process (Croskerry, 2009). For example, Alter, Oppenheimer, Epley and Eyre (2007) suggested that task difficulty levels guide the subsequent use of either the intuitive System 1 or analytical System 2. Thus, the current study also examines the relationships between metacognitive awareness, case difficulty, and decision-making style.

This work builds on an existing dataset of image-based clinical reasoning, in which speech data were collected from physicians as they inspected images of dermatological cases and moved towards diagnosis (Hochberg, Alm, Rantanen, DeLong, & Haake, 2014a). Physician narratives were annotated for decision style, and these annotations were used as class labels to build a model to automatically classify decision style based on linguistic, diagnostic difficulty, and demographic features. Decision annotation metrics were also used to investigate the relationships between decision style, expertise, accuracy, and metacognitive awareness.

The study rests on several key assumptions about the nature of the speech data collected from clinicians engaged in a modified Master-Apprentice task (Beyer & Holtzblatt, 1997), in which they describe the diagnosis as if teaching a student. First, it assumes that their verbalizations, while also partially reflective of a teaching process, are similar to think-aloud data, as both involve real-time verbal descriptions of a task as participants perform it. If so, the claim made of think aloud data – that they reflect the contents of working memory in a relatively unaltered form (Ericsson & Simon, 1993) – may also be made for the data analyzed in the current work. Second, it assumes that the linguistic data reflect underlying cognitive processes, and that individual differences in linguistic style are of interest in studying these processes (Pennebaker & King, 1999; Cohn, Mehl, & Pennebaker, 2004).

The next section will discuss each variable in detail through the lens of the existing literature. The methods section will describe the data collection process, performed as part of the Human-Centered Computing Group at the Rochester Institute of Technology (see Womack, Alm, Calvelli, Pelz, Shi, & Haake, 2013), as well as the corpus annotation for decision style. Finally, the results of the computational model of decision-making style, as well as statistical analyses on decision style, metacognition, and related variables, are discussed.

Literature Review

What is metacognition?. Metacognition is a fuzzy concept; there is not one universally accepted definition (see Dinsmore, Alexander & Loughlin, 2008, for a detailed review). However, most attribute the construct's beginnings to a now-classic article by Flavell (1979), who defined metacognition as "knowledge and cognition about cognitive phenomena" (p. 906). Flavell proposed two types of metacognition: metacognitive knowledge and metacognitive experiences, where metacognitive knowledge concerns “stored world knowledge of...cognitive tasks, goals, actions and experiences”, while metacognitive experiences are “cognitive or affective experiences” relevant to any intellectual activity (p. 906).

Flavell (1979) further divided metacognitive knowledge into knowledge of a person, task, or strategy. Metacognitive knowledge of a person concerns knowledge of individuals as cognitive processors, while metacognitive knowledge of a task concerns its demands, difficulty, the information available to complete it, and the chance of success in meeting the desired goal. Finally, metacognitive knowledge of strategy is concerned with determining and monitoring the best cognitive strategies for achieving a certain goal. Metacognitive experiences, in turn, are likely to occur in situations that engage highly involved, analytical thinking or new situations; they, in turn, can contribute to metacognitive knowledge (Flavell, 1979). For example, the feeling that one is not using the best cognitive strategy to achieve a particular task can then influence one’s knowledge and beliefs of what strategies are appropriate in subsequent similar tasks (see Garofalo & Lester, 1985, for a review of interactions between components of metacognitive knowledge).

Since Flavell’s work, most researchers have split metacognition into two main components that do not necessarily align with Flavell’s distinctions. Almost all agree on the existence of a metacognitive knowledge component, while the second component generally concerns metacognitive regulation, or what would be equivalent to Flavell’s metacognitive strategies component of metacognitive knowledge. This two-component model was first outlined by Baker and Brown (1984).

Thus metacognition concerns an individual’s knowledge of cognition in themselves and others, of cognitive strategies, and of the interaction between task demands and cognition; it also concerns an individual’s effective use of this knowledge base to plan, monitor, evaluate, and update cognitive processes. The first term, metacognitive knowledge, has also been called self-appraisal of cognition (Paris & Winograd, 1990), as well as declarative knowledge (Kluwe, 1982). The second term, metacognitive regulation, has also been termed self-management of cognition (Paris & Winograd, 1990), executive processes (Kluwe, 1982), and metacognitive skills (Brown, 1978). In an exhaustive review of the origins of metacognition in academic research, Hacker (1998) nicely defines metacognitive knowledge as “knowledge of one’s knowledge, processes, and cognitive and affective states” and metacognitive regulation as “the ability to consciously and deliberately monitor and regulate one’s knowledge, processes, and cognitive and affective states” (para. 27).

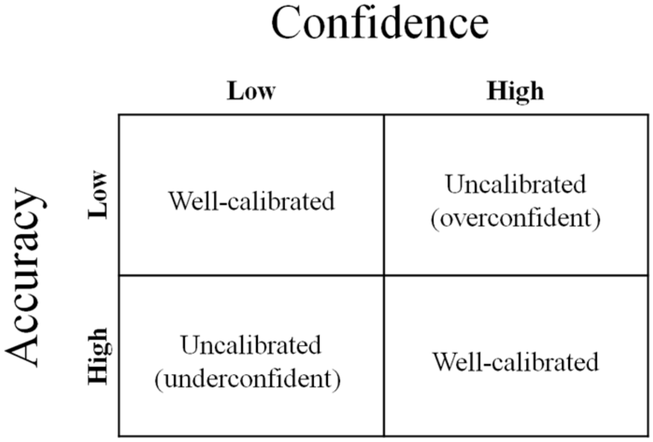
Research on metacognition. During the first few decades of work on metacognition, there were three main categories of research: studies of cognitive monitoring, studies of cognitive regulation in response to changing task demands, and studies of both monitoring and regulation. The first, cognitive monitoring, included studies of the tip of the tongue phenomenon, allocation of study effort, and judgments of learning. In studies of cognitive regulation, participants first perform a training task, and are then tested on strategy used in a similar task; this line of study is concerned with how and to what extent individuals determine which strategies are most effective for a particular task. Finally, studies of both monitoring and regulation are mostly studies of memory and concern the use of particular strategies to facilitate recall (Hacker, 1998).

In the last decade, a new area of study has emerged: the study of metacognition in educational contexts. This line of study attempts to take theoretical knowledge of metacognition and put it to use in order to improve learning and teaching. Most studies examine whether metacognitive theory can improve learning, and the overwhelming majority of studies find that indeed it can (see Hacker, 1998, for a review). Studies of metacognition in education have examined the domains of reading comprehension (Artz & Armour-Thomas,1992; Mokhtari & Reichard, 2002), mathematical skill (Garofalo & Lester, 1985), physics problems (Neto & Valente, 1997), and medical education, both at the theoretical (Croskerry, 2003a, 2003b; Croskerry & Norman, 2008) and empirical (Coderre, Wright, & McLaughlin, 2010; Mamede, van Gog, van den Berge, Rikers, van Saase, van Guldener, & Schmidt, 2010; Sherbino, Dore, Siu, & Norman, 2011) levels. Importantly, researchers on metacognition in education note that metacognitive skills should not be considered an end, but rather a means to an end (Paris & Winograd, 1990). That is, it is not enough to teach metacognition as educational content — a static set of facts. Instead, students should be taught how to incorporate metacognitive strategies into learning in a variety of domains.

Measurement of metacognition. In their comprehensive review on the assessment of metacognition, Pintrich, Wolters, and Baxter (2000) suggest a three-component model of metacognition, as opposed to the two-component model more commonly found in the literature. Rather than just a metacognitive knowledge and metacognitive monitoring and control components, the authors suggest that monitoring and control be split into two separate components. *Metacognitive awareness and monitoring* refers to online (real-time) monitoring and assessment of task performance, while the self-regulation and control component actively implements changes to cognitive processes, based on knowledge gained from both metacognitive knowledge (a static, relatively stable component), and metacognitive awareness and monitoring (a dynamic component). *Self-regulation and control* includes planning, strategy selection and use, resource allocation, and volitional control, or control of motivation, emotions and the environment (Pintrich et al., 2000). The current work adopts this framework to frame, theoretically, research on the assessment of metacognition, and focuses particularly on *metacognitive awareness*, consistently with the literature treating this component as distinct and empirically measurable (Schraw, 2009). However, as Pintrich et al. themselves note, researchers have found it difficult to disentangle, and measure separately, monitoring and control processes in empirical research.

Pintrich et al. (2000) discuss four measures of metacognitive awareness and monitoring: ease-of-learning judgments (EOL), judgments-of-learning (JOL), feeling of knowing (FOK) judgments, and confidence ratings. Ease of learning judgments occur prior to a task. When first engaging in a new task, individuals make judgments with respect to the difficulty level of a learning task. Judgments of learning occur during or after the task, and are concerned with online monitoring. An individual might assess the extent to which they understand or do not understand specific parts of a learning task. Next, feelings of knowing, which include tip-of-the-tongue (TOT) phenomena, correspond to judgments that an individual knows they know something, but cannot access or recall the knowledge at that moment

(Pintrich et al., 2000). Finally, confidence judgments occur either during a task (concurrent judgments) or after a task has been completed (retrospective judgments; see Schraw, 2009, for a review). Individuals rate their confidence in the level of their performance, generally on a scale from 0 to 100 (Pintrich et al., 2000). It is confidence judgment, with respect to diagnostic accuracy, which is the main measure of metacognitive awareness in the current study. Individuals whose confidence judgments are good predictors of their performance are said to be *well-calibrated* (Pintrich et al., 2000; and see Figure 1, below). Alternatively, such individuals may also be considered as *possessing higher levels of metacognitive awareness.*



		Confidence	
		Low	High
Accuracy	Low	Well-calibrated	Uncalibrated (overconfident)
	High	Uncalibrated (underconfident)	Well-calibrated

Figure 1. A schematic representation of the relationships between confidence and accuracy.

Confidence judgments as measures of metacognitive awareness. This section discusses the last category, confidence judgments, in greater detail, as they are the primary indicators of metacognitive awareness in the current study. Confidence judgments are evaluated with respect to task accuracy, and there are five widely-used measures of confidence-accuracy relationship. *Absolute measures* compare confidence for an item or set of items to accuracy as measured by the proportion of items correct for that item or set of items, while *relative measures* represent the extent to which higher confidence is associated with higher accuracy (Baranski & Petrusic, 1998; Krug, 2007; Nelson, 1996; Weber & Brewer, 2003). The current study employed relative measures of the confidence-accuracy relationship, as an important fraction of study participants did not provide confidence judgments towards the lower end of the scale. This may suggest high confidence, but it may also suggest that individual physicians simply adjusted the scale so that 40-50 percent confidence was representative, for them, of low confidence. To address this issue, some studies employ half-range confidence scales, ranging from 50 to 100 percent. Further, scholars have suggested that absolute (rather than relative) measures are most appropriate for tasks employing half-range confidence scales (Harvey, 1997; Weber & Brewer, 2003). There are two commonly used relative measures of the confidence-accuracy relationship. One is the point-biserial correlation (r_{pb}), which is a correlation appropriate for ordinal and interval or ratio data; it is mathematically equivalent to the Pearson statistic. The point-biserial correlation formula is computed as follows:

$$r_{pb} = \frac{M_p - M_q}{S_t}\sqrt{pq} \tag{1}$$

where M_p is the mean for the interval variable values for which the ordinal value is coded 1, M_q is the mean for the interval variable values for which the ordinal value is coded 0, S_t is the standard deviation for the interval variable, p is the proportion of interval variable values coded 1, and q is the proportion of interval variable values coded 0.

The other is the nonparametric Goodman-Kruskal gamma correlation, which is a rank correlation used often in the psychological literature, including to examine the confidence-accuracy relationship. In this specific application, it considers whether confidence is higher for correct than incorrect responses, and vice versa. Like the Pearson statistic, it also ranges from -1 to +1 (Bornstein & Zickafoose, 1999; Krug, 2007; Nelson, 1984). The formula for the gamma correlation is as follows:

$$G = \frac{N_a - N_i}{N_a + N_i} \tag{2}$$

where Na is the number of aligned pairs and Ni is the number of inverted pairs.

With respect to the current work, the relationship between physicians' confidence and accuracy is of interest because participants with high levels of metacognitive awareness and monitoring are likely to be well-calibrated. Particularly, poor calibration may negatively impact cognitive regulation. As per Pintrich et al. (2000), "if the students believe that they are learning, when they are not, then they will be unlikely to change or effectively regulate their cognition and learning" (p. 90).

Methodology for measurement of metacognitive awareness. In terms of methodology, the measurement of metacognitive awareness and regulation falls into four main categories: self-report, error-detection studies, interviews, and think-aloud protocols (Pintrich et al., 2000; Dinsmore, Alexander, & Loughlin, 2008). In studies of monitoring, self-report judgments include ease of learning judgments, judgments of learning, feelings of knowing, and confidence judgments. Measures of these judgments are then compared with actual task performance. In studies of regulation, self-report questionnaires include the Learning and Study Strategies Inventory (LASSI; Weinstein, Schulte, & Palmer, 1987) and the Motivated Strategies for Learning Questionnaire (MLSQ; Pintrich & DeGroot, 1990). In error-detection studies, participants are asked to read a text and report any contradictions, omissions, or other errors. Students who notice more errors are considered better monitors of reading comprehension. Finally, interviews, both formal and informal (Artz & Armour-Thomas, 1992; Swanson, 1990; Zimmerman & Martinez-Pons, 1986) and think-aloud protocols both offer rich verbalizations, descriptions, and explanations, which are then studied for evidence of monitoring and regulation, via methods such as protocol analysis (Pintrich et al., 2000).

Studies of verbal data in clinical reasoning. In this work, spoken descriptions were elicited from physicians using a modified Master-Apprentice scenario, described in more detail in the methods section. These data are similar in nature to the speech data collected in think-aloud studies, in which participants describe the task as they go about it (Duncker & Lees, 1945; see Nielsen, Clemmensen, & Yssing, 2002, for a review). Specifically, the think-aloud methodology has been promoted as a fitting tool for the examination of clinical reasoning (Lundgrén-Laine & Salanterä, 2010).

Lundgrén-Laine and Salanterä note that think-aloud studies allow for linking thought processes with concurrent perceptions, and that the data collected are particularly rich, detailed, and fine-grained. Particularly, they advocate the use of protocol analysis to examine the resulting data of think-aloud studies. For example, Backlund, Skånér, Montgomery, Bring, and Strender (2003) asked 20 doctors to think aloud and describe their diagnoses and suggest treatment in response to six text-based case descriptions of patients with elevated cholesterol. Between subsequent revelations of case information (e.g. demographic information, lab results), they were also asked to rate their likelihood of prescribing pharmacological treatment. Participant transcripts were divided into statements and classified as one of ten categories. Categories included *attention* (repeating or reading basic information), *evaluation* (considering information with respect to treatment), *rule* (general domain-specific principles), and *explanation* (inferences). Backlund et al. (2003) also concluded that think-aloud data, in conjunction with protocol analysis, are effective in studying clinical reasoning, as participants' verbalizations were consistent with their on-line ratings of whether they were likely to describe medication.

Interestingly, one study of general problem-solving, which had participants solve the Tower of Hanoi task, has even linked think-aloud verbalization to the promotion of metacognitive monitoring and regulation (Berardi-Coletta, Buyer, Dominowski, & Rellinger, 1995). In medicine, think-aloud seminars have been used to teach clinical reasoning skills, though there are not yet uniform guidelines for doing so (see Banning, 2008, for a review).

However, there are several caveats that must be taken into account when employing think-aloud tasks. First, verbalization may add cognitive load and impact the allocation of attention, thus influencing task performance. The extent to which this is likely to occur is known to be affected by participant age, difficulty of primary task, and verbal ability (Ericsson & Simon, 1980; Pintrich et al., 2000). In addition, the task instructions are likely to influence the nature of the verbalizations. Notably, a study of 55 radiographers who described their impressions of videotaped clinical scenarios found that their decision-making processes were relatively unstructured (Prime & Le Masurier, 2000). This is in contrast to a recent study of physician decision-making in dermatology, in which most physicians generally followed this pattern: symptoms and morphology, differential diagnosis, final diagnosis (McCoy et al., 2012); such structure may be reflective of their training.

Instructions for think-aloud verbalizations may vary in their effects on individuals. For example, any additional cognitive load caused by verbalization may be more detrimental to novice participants (residents) than expert participants (attendings), since experts are able to increase working memory capacity by employing schemas developed via expertise (Kalyuga, Ayres, Chandler, & Sweller, 2003; Paas, Renkl, & Sweller, 2003). However, there is some evidence supporting that verbal protocol studies change the nature of cognitive processes only in the case of Level 3 verbalizations (Ericsson & Simon, 1993), in which participants must reflect when prompted (see Bannert & Mengelkamp, 2008, for a review).

Decision-making style. The discussion will now turn to the second variable, decision-making style, describing both the general theory and its application in clinical contexts. Klein (1999) has proposed that individuals making decisions in their domain of expertise often utilize *recognition-primed decision-making* based on automatic processing and retrieval of past knowledge. This type of decision-making appears to rely primarily on System 1 processes, as discussed in dual-process models of cognition (see Evans, 2008, for a review). System 1 is characterized by its automatic, rapid, and unconscious nature; System 2, in contrast, is a controlled, slow, and conscious mode of thought (Evans, 2003; Kahneman & Frederick, 2002; Stanovich & West, 2000).

In fact, Evans specifically makes a distinction between heuristic (System 1) and analytic (System 2) modes of reasoning (Evans, 1989, 2006). It is this label, heuristic, that highlights the use of rules of thumb, or mental shortcuts, in this mode of decision-making. Use of the heuristic system, while often efficient and useful, may lead to cognitive errors based on heuristics and biases, generally (Shanteau, 1988) as well as more specifically in the medical domain (Croskerry, 2003b; Graber, 2009).

Since individuals tend to rely more on heuristic reasoning under time pressure (Rieskamp & Hoffrage, 2008; Evans & Curtis-Holmes, 2005) or in the face of uncertainty (Hall, 2002), which occur often in medicine, they are more liable to make cognitive errors due to biases and heuristics in such conditions. Croskerry (2003b) classified over 30 such biases and heuristics, or, cognitive dispositions of respond (CDRs) that underlie diagnostic error in medicine, including: anchoring, base-rate neglect, the framing effect, hindsight bias, gambler's fallacy, premature closure, representativeness.[2] In fact, Berner and Graber (2008) note that diagnostic error in medicine is estimated to occur at a rate of 5-15%, and that two-thirds of diagnostic errors involve cognitive root causes.[3]

Thus, the current study attempts to distinguish between intuitive and analytical decision-making processes in clinician verbalization, based on linguistic, demographic, and case difficulty features. To operationalize decision-making, one of the two main theoretical models must be chosen as a basis for prediction. The current study employs the dual process decision-making framework (as in Evans, 2003), for two reasons. First, the recognition-primed decision-making model focuses on expert decision-making, while dual process theory applies to the full gamut of decision makers, from novice to expert. In this work, the wide range of professional experience and training level among the physician participants, ranging from several years to several decades, means that the dual process framework is more appropriate than the recognition-primed decision-making model. The second reason concerns quantification: the ease with which variables in each theory are naturally operationalized. The intuitive-analytical spectrum, following Hammond's cognitive continuum (discussed below), is one that is particularly conducive to the evaluation of decision-making style via a rating scale.

The Cognitive Continuum framework. Hammond (1981) developed the Cognitive Continuum Theory to describe the relationship between tasks and modes of cognition. In this framework, *intuitive* reasoning is described as rapid, unconscious, moderately accurate, employing simultaneous use of cues, and involving pattern recognition (Hammond, 1981). *Analytical* decision-making is described as slow, conscious, more accurate, making sequential use of cues, based on logical rules, and task-specific (Hammond, 1996). According to

[2] Such biases have been reported across domains.

[3] The two additional major error categories in medicine include *system* and no-fault errors (see Graber, Gordon & Franklin (2002) for a review).

Hammond's theory, however, most reasoning occurs between the two poles of purely intuitive and purely analytical decision-making; this type of reasoning is known as *quasirational* reasoning (Hammond, 1981; Hamm, 1988). Quasirationality may be characterized by a mix of, or oscillation between, intuitive and analytical reasoning, or by intermediate values of the features that define decision-making along the intuitive-analytical continuum.

Hammond suggested that certain tasks best fit certain modes of cognition (Hammond, 2000). Hamm (1988), who wrote extensively on Cognitive Continuum Theory, reviews three main task features, attributes of which induce either analytical or intuitive decision-making. Task features include task structure, task ambiguity, and task presentation. With respect to task structure, well-structured tasks encourage analytical decision-making, while ill-structured tasks induce intuitive decision-making. With respect to task ambiguity, organizing principles and unfamiliarity with task content are likely to induce analytical thinking. Finally, with respect to task presentation, if tasks are decomposed into subtasks, analytical reasoning is induced. Also, if stimuli are presented pictorially rather than quantitatively, and if individuals work under time pressure, they are more likely to use intuitive modes of cognition (Hamm, 1998).

Cader, Campbell and Watson (2005) suggest that cognitive continuum theory is appropriate for evaluation of decision-making in nursing and medical contexts. They praise its parsimony and testability: the ease of operationalizing the framework in empirical research. They further claim that the representation of intuitive and analytical decision making on a continuum, rather than as a dichotomy, is a particular strength of the theory, as nurses (Cader et al.'s population of interest) often use a quasi-rational mode of cognition.

Lauri et al. (2001) collected data from 459 nurses in five countries, who completed a 56-item domain-specific questionnaire corresponding to the four main stages of decision-making in nursing: collecting information; processing information; planning; and implementing, monitoring, and evaluation. Half of the questionnaire items referred to intuitive processes, and half to analytical processes. For example, participants rated the following statements on a 5-point Likert scale from always to never: *I draw on nursing process thinking to define* *the patient's nursing problem* (analytical) or *It is easy for me to see, even without closer analysis, which pieces of information are relevant to defining the patient's nursing problems* (intuitive). Factor analysis revealed five factors, termed models, of decision-making: one intuitive, one analytical, and three quasi-rational. Models used varied by country and whether nurses were in short or long term care. There are at least two study limitations, however: questionnaires were translated since they were administered in more than one country, and the study did not employ random sampling. Nonetheless, the study suggests that different modes of reasoning are used in different nursing contexts, and that the predominant modes are quasirational, in the middle of the cognitive continuum (Lauri et al., 2001).

The current study extends the application of Cognitive Continuum Theory from the study of nursing decision-making to the study of physician decision-making in the domain of dermatology. Decision-making, as it appears in physician verbalizations, was evaluated with respect to four zones on the cognitive continuum: intuitive decision-making, analytical decision-making, and two intermediate regions, representing two quasirational modes of decision-making, one closer to the intuitive end of the scale and the other closer to the analytical end of the scale.

Purpose of the Research

The purpose of this research is to test the following thesis: *Decision making style can be reliably annotated for narratives of diagnostic reasoning. Furthermore, linguistic and other features associated with the narratives allow for automatic annotation of decision style.*

This thesis rests on two key assumptions: first, that verbal data reflect the contents of working memory (Ericsson & Simon, 1993), and second, that cognitive processes are revealed in language use (Pennebaker & King, 1999; Cohn, Mehl, & Pennebaker, 2004). However, it is reasonable to assume that some, albeit not all, cognitive processes may be revealed in language use, since not every part of a decision process is available to consciousness (see Bannert & Mengelkamp, 2008, for a review).

Two more assumptions concern the validity of measures used in the study, including related hypotheses on decision-making, expertise, and metacognitive awareness: first, that metacognitive awareness can be measured via the correlation between confidence and accuracy (Pintrich, Wolters, & Baxter, 2000), and second, that physicians' diagnostic narratives can be coded on an intuitive-analytical continuum of decision-making style (Hammond, 1981; Lauri & Salanterä, 1994).

Hypotheses. This section lays out the first hypothesis, which corresponds to the thesis on the manual and automatic annotation of decision style in physician verbalizations. Six other hypotheses concern the relationships between decision style, expertise, metacognitive awareness, and case difficulty, and diagnostic accuracy. (For a diagram illustrating the hypotheses and the major study variables, see Appendix A.)

H1. Decision making style can be reliably annotated for narratives of diagnostic reasoning. Furthermore, linguistic and other features associated with the narratives allow for automatic annotation of decision style.

Further, it has been suggested that intuitive reasoning relies on heuristics and biases, which are often sources of error in clinical reasoning (Croskerry, 2003b; Graber, 2009). Therefore:

H2. Intuitive reasoning will be associated with lower levels of diagnostic accuracy.

In addition, experts have more experience, including a broad base to drawn upon in the use of pattern recognition underlying intuitive decision-making (Klein, 1999; Croskerry, 2006). Thus:

H3. Experts will have better success with intuitive reasoning than novices.

Moreover, previous work has linked perceived difficulty to increased use of analytical decision-making (Alter, Oppenheimer, Epley, and Eyre, 2007). Accordingly:

H4. More difficult cases will be associated with analytical decision-making, while less difficult cases will be associated with intuitive decision-making.

Metacognitive training has also been shown to aid decision-making (Batha & Caroll, 2007). In addition, in studies of reading (Paris & Oka, 1986) and mathematics knowledge (Tobias & Everson, 1995), metacognitive awareness has been linked to higher performance.

Thus:

H5. Higher levels of metacognitive awareness will be associated with higher levels of diagnostic accuracy.

In addition, metacognition has been deemed key to the development of expertise in general (Sternberg, 1998) and in medicine (Quirk, 2006). Therefore:

H6. Experienced physicians will exhibit higher levels of metacognitive awareness than inexperienced physicians.

The final hypothesis concerns the link between metacognition and decision-making style. As noted above, metacognitive experiences of disfluency/difficulty have been shown to serve as cue for analytical decision-making (Alter, Oppenheimer and Eyre, 2007; Thompson, 2009). Further, metacognitive awareness prompts individuals to switch to the analytical System 2 when necessary (Croskerry, 2009). Thus, since physicians with higher levels of metacognitive awareness may be more attuned to switch-inducing disfluency cues:

H7. Higher levels of metacognitive awareness will be associated with increased use of analytical decision-making.

Contributions. The current work will shed light on the links between decision style, metacognitive awareness, expertise, and diagnostic difficulty. This work will also add to the small base of literature that has reported studies of the link between metacognition and decision-making, providing more information on the association between the two. In addition, the study offers a methodological contribution with respect to the annotation scheme developed for corpus annotation of decision style.

With respect to decision style modeling, this appears to be the first study attempting to computationally predict physician decision style. Similar to the case of affect (Alm, 2011), automatic annotation of decision style can be characterized as a subjective natural language processing problem. This adds special challenges to the modeling process. Accordingly, this work details a thorough process for moving from manual to automatic annotation.

This study contributes to annotation methodology, cognitive psychology, and clinical computational linguistic analysis. Methodologically, the study details a careful process for selecting and labeling manually annotated data for modeling in the realm of subjective natural language phenomena, thus addressing the need for their characterization (Alm, 2011). Theoretically, acceptable annotator reliability on decision style, along with successful computational modeling, will lend construct validity to the dual process model. From a linguistic perspective, the identification of discriminative features for intuitive and analytical reasoning provides a springboard for further studying decision-making using language as a cognitive sensor.

Practically, prediction of decision style would also be useful for determining whether individuals are using the appropriate style for a particular task, based on analyses linking decision style to task performance. Thus, in the case of successful modeling of decision style, this work will provide preliminary support for the development of a new linguistic measure of decision-making style, which may be derived in real-time. Importantly, detection of decision style from observable linguistic behaviors allows for objective measurement that avoids biases present in self-report surveys (Sjöberg, 2003; Allinson & Hayes, 1996).

Method

The current work makes makes secondary use of one dataset from existing resources, originally collected to explore research questions concerning image-based diagnostic reasoning (Li, Pelz, Shi, & Haake, 2012; Womack, Alm, Calvelli, Pelz, Shi, & Haake, 2013). Eye-tracking and speech data were collected from dermatologists as they described their evaluation of images presenting different dermatological diagnoses, with the intent of characterizing perceptual and conceptual components of image understanding and infusing expertise into the design of an image-retrieval system. Physician eye-tracking patterns, and their verbalizations, could then be used to develop semantic metadata and discover new relationships among images, to inform a novel image retrieval system design. Thus, such elicited data served to gather elements of domain knowledge and cognitive processing that could be used to evaluate human image understanding and incorporated into the storage of and search for images (Guo, Li, Alm, Yu, Pelz, Shi, & Haake, 2014).

Participants

Participants were physicians (*N*= 29; 16 women, 13 men) attending a U.S.-based dermatology conference in 2011. Eleven were board-certified dermatologists (8 of whom were educators) and 18 were resident physician dermatologists in training. Participants' institution of medical training, years of experience, and whether they were educators was recorded; age was not recorded. Participants hailed from over nine institutions, and their experience ranged from several months to 38 years. Two participants were not native English speakers. The participants received $25 in compensation for their participation in the study and had a chance of winning an iPad.

Apparatus and Stimuli

Participants were shown 30 images of dermatological symptoms (see Figure 2, below), indicative of a range of dermatological conditions, on a computer monitor. Images represented both common (e.g., seborrheic dermatitis) and rare (e.g., lymphomatosis papulosis) conditions, and varied in diagnostic difficulty. The images were the sole source of information available to the participants (i.e., no patient histories or demographic details were made available). In this work, each image represents a case; henceforth, each of these stimuli will be referred to interchangeably with the terms *image*, *case*, or *image case*.

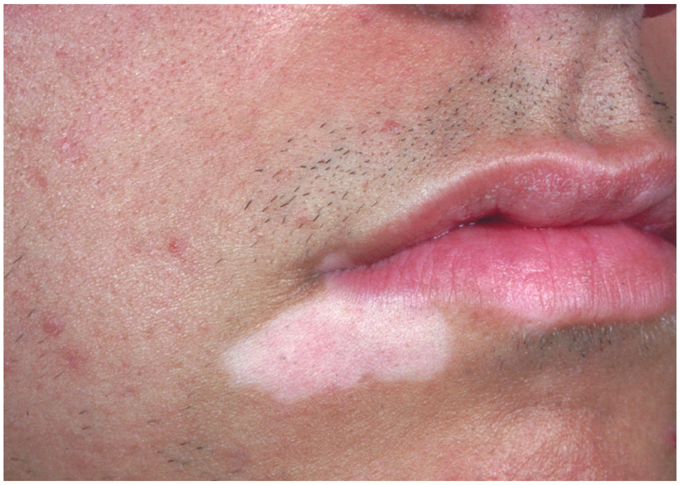


Figure 2. Example of image case shown to participants, used with permission from Logical Images, Inc.

Procedure

Each participant saw all 30 images in random order. The participants were asked to discuss each case as if teaching a student, in a modified Master-Apprentice scenario (Beyer & Holtzblatt, 1997). The Master-Apprentice scenario is traditionally used in human-computer interaction studies, in which a researcher serves as the apprentice and learns from an expert in some task; the expert may be a professional or even a long-time customer. The master-apprentice scenario has several benefits. First, it assumes that various subtasks, and the reasons for performing each subtask, are most available to the master as they perform the task itself. Second, the master-apprentice scenario is a teaching scenario, so it encourages experts to provide detailed descriptions, with the goal of promoting understanding in the apprentice. Finally, it promotes a more complete description of the task and the role of various contextual details, as many experts use environmental cues as triggers to action (Beyer & Holtzblatt, 1997). In the present study, the scenario was modified in that no apprentice or student was actually present.

Participants were asked to describe the case presented to them, and to suggest a differential diagnosis (a range of potential diagnoses) and then a final diagnosis. They also provided a confidence/certainty score, on a scale of 0-100. Along with the audio recordings, eye movements were tracked and recorded.

There was no time constraint or total time estimate given to participants. Participants received information at the study outset was that it would include a total of 30 images. Two students were present in the room, in charge of the eye movement and audio data collection.

A total of 868 narratives (29 participants X 30 images, minus 2 images skipped during the procedure) were included in this study's dataset version. Each narrative is about one minute in length, for a total of 15.8 hours of audio recordings.

Preliminary Data Analysis

Narratives were transcribed and time-aligned to the recordings. A licensed dermatologist on the research team evaluated each narrative for accuracy on the basis of the final diagnosis. Narratives received one of five scores: *incorrect* (57% of narratives), *correct* (39%), *not given* (1%), *half* (in the case that two final diagnoses were provided and only one was correct; 1%), and *partial* (in the case that the correct pathological category was noted, but not the specific diagnosis; 2%).

For the purposes of this work, the *half* score was counted as *correct*, while the *partial* score was counted as *incorrect*. This reasoning is in line with other work on this dataset (e.g., Bullard, Alm, Qi, Shi, & Haake, 2014), and based on the logic that in the case of a *half* score, the physician did in fact identify the diagnosis and would be likely to determine the correct diagnosis with follow-up treatment, while physicians whose narratives were scored *partial* did not clearly indicate or identify the specific correct diagnosis. In addition, in this work, the terms *correctness* and *accuracy* are used interchangeably to refer to diagnostic accuracy, based on the accuracy of final diagnosis as discussed here.[4]

[4] While *accuracy* is the term used in the literature on metacognitive awareness, previous published work on this research study, by multiple authors, has employed the term *correctness* (e.g., Womack, Alm, Calvelli,

Corpus Annotation of Decision Style

The corpus of physician narratives was annotated for decision style in a pilot study and then a main annotation study (Figure 3).[5] Two male annotators with graduate training in cognitive psychology independently rated each narrative on a 4-point scale from *intuitive* to *analytical* (Figure 4). The two middle labels reflect the presence of both styles, with intuitive (*BI*) or analytical (*BA*) reasoning being more prominent.[6]

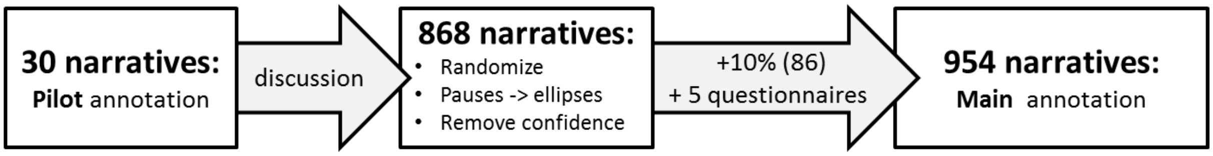


Figure 3. Overview of annotation methodology. Conclusions from the pilot study enhanced the main annotation study. To ensure high-quality annotation, narratives appeared in random order, and 10% (86) of narratives were duplicated and evenly distributed in the annotation data, to later assess intra-annotator reliability. Questionnaires were also interspersed at 5 equal intervals to study annotator strategy.

Narratives were presented to annotators as anonymized transcripts, and were not accompanied by any additional information. Annotator instructions included a definition and characteristics of each decision style, as well as examples of narratives corresponding to each of the four decision ratings (see Appendix B). Since analytical reasoning involves detailed examination of alternatives, annotators were asked to avoid using length as a proxy for decision style.

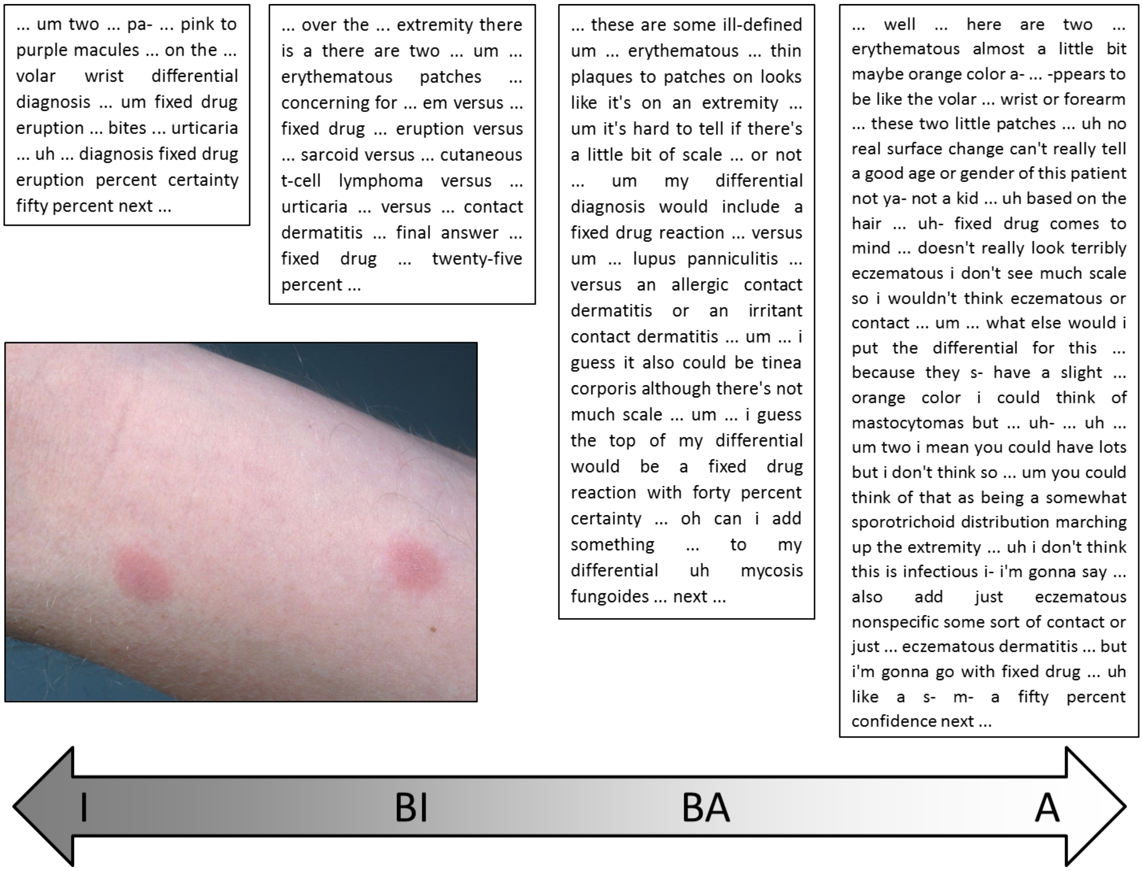
Several measures were taken to ensure high-quality annotation of the narratives. The order of the narratives was randomized. Then, to measure intra-annotator reliability, 10% (86) of the narratives were duplicated and added to the annotation data; each duplicate appeared 221 narratives after its first occurrence.[7] Finally, five questionnaires were evenly

Pelz, Shi, & Haake, 2013; Bullard et al., 2014).

[5] Within a reasonable time frame, the annotations will be made publicly available as part of a corpus release.

[6] As noted in the annotator instructions, these middle categories could reflect intermediate values of the features distinguishing decision styles; or a mix of characteristics, from both intuitive or analytical modes; or reasoning that oscillates between the two modes.

[7] To reduce the possibility that annotators would recognize a duplicated narrative; the +221 count for later narratives was wrapped around to the start.



... um two ... pa- ... pink to purple macules ... on the ... volar wrist differential diagnosis ... um fixed drug eruption ... bites ... urticaria ... uh ... diagnosis fixed drug eruption percent certainty fifty percent next ...

... over the ... extremity there is a there are two ... um ... erythematous patches ... concerning for ... em versus ... fixed drug ... eruption versus ... sarcoid versus ... cutaneous t-cell lymphoma versus ... urticaria ... versus ... contact dermatitis ... final answer ... fixed drug ... twenty-five percent ...

... these are some ill-defined um ... erythematous ... thin plaques to patches on looks like it's on an extremity ... um it's hard to tell if there's a little bit of scale ... or not ... um my differential diagnosis would include a fixed drug reaction ... versus um ... lupus panniculitis ... versus an allergic contact dermatitis or an irritant contact dermatitis ... um ... i guess it also could be tinea corporis although there's not much scale ... um ... i guess the top of my differential would be a fixed drug reaction with forty percent certainty ... oh can i add something ... to my differential uh mycosis fungoides ... next ...

... well ... here are two ... erythematous almost a little bit maybe orange color a- ... -ppears to be like the volar ... wrist or forearm ... these two little patches ... uh no real surface change can't really tell a good age or gender of this patient not ya- not a kid ... uh based on the hair ... uh- fixed drug comes to mind ... doesn't really look terribly eczematous i don't see much scale so i wouldn't think eczematous or contact ... um ... what else would i put the differential for this ... because they s- have a slight ... orange color i could think of mastocytomas but ... uh- ... uh ... um two i mean you could have lots but i don't think so ... um you could think of that as being a somewhat sporotrichoid distribution marching up the extremity ... uh i don't think this is infectious i- i'm gonna say ... also add just eczematous nonspecific some sort of contact or just ... eczematous dermatitis ... but i'm gonna go with fixed drug ... uh like a s- m- a fifty percent confidence next ...

Figure 4. Four narratives along the intuitive-analytical decision-making continuum, for which annotators agreed on their labels, where *I=Intuitive, BI=Both-Intuitive, BA=Both-Analytical, A=Analytical.* The narratives were produced by different physicians for the same image case (left, used with permission from Logical Images, Inc.), and all four physicians were correct in their final diagnosis. (Confidence mentions were removed in narratives presented to annotators, to avoid any potential bias.)

spaced among the narratives, which surveyed annotators for their comments and queried them as to the relative importance of various factors (e.g., silent pauses, or use of justification) in annotation. In data analysis, primary ratings (the first time the annotators encountered each narrative) were used for descriptive statistics and inter-annotator reliability, while secondary ratings were used to determine intra-annotator reliability.

After the pilot, the annotators jointly discussed disagreements with one researcher. Inter-annotator reliability, measured by linear weighted kappa (Cohen, 1968), was 0.4 before and 0.8 after resolution.

Both annotators reported using the physician-provided confidence as a factor in determining the decision-making style of the narrative; analysis of the 30 pilot ratings confirmed this trend. Thus, in subsequent annotation confidence mentions were removed if they appeared after the final diagnosis (most narratives), or, if interspersed with the diagnostic reasoning, replaced with dashes (10% of narratives). For example, *eighty percent sure this is a case of contact dermatitis* would be changed to — *percent sure this is a case of contact dermatitis*.[8] Finally, silent pauses[9] were coded as ellipses to aid in the human parsing of the narratives.

[8] In addition to specific numbers (e.g., *ninety*), physicians also expressed their confidence with quantifiers (e..g, low, high) and, rarely, with direct statements (e.g., *I am not certain*); all were replaced with dashes. This measure ensured that annotations corresponded to decision style and not confidence.

[9] As based on provided transcripts (above around 0.3 seconds; see Lövgren & Doorn, 2005).

Results

Decision Style Annotation

This section details the results of the corpus annotation of decision style. The distribution of annotator ratings, annotator reliability, annotator strategy, and a narrative case study are discussed.

Quantitative annotation analysis. Table 1 shows the annotator rating distributions on the 4-point decision rating scale.[10] Though Annotator 1's ratings skew slightly more analytical than Annotator 2, a Kolmogorov-Smirnov test showed no significant difference between the two distributions ($p = 0.77$).

Table 1

Distribution of Annotator Ratings

	Rating			
Annotator	I	BI	BA	A
Annotator 1	89	314	340	124
Annotator 2	149	329	262	127

N = 867. *I = Intuitive, BI = Both-Intuitive, BA = Both-Analytical, A = Analytical.*

Figure 5 shows visually the distribution of annotation labels for both annotators, respectively, for the whole dataset, on the original 4-point scale.[11] In comparison, Figure 6 shows the annotators' distributions across a collapsed 2-point scale of intuitive vs. analytical, where, for each annotator, narratives labeled *BI* were assigned to *I* and those labeled *BA* assigned to *A*. A 2-point collapsed scale was used for the purposes of computational modeling of decision style (discussed below).

Annotator agreement was well above chance for both the 4-point (Figure 7) and 2-point (Figure 8) scales. Notably, the annotators were in full agreement or agreed within one rating for over 90% of narratives on the original 4-point scale. This pattern of variation reveals both

[10] N = 867 after excluding a narrative that, during annotation, was deemed too brief for decision style labeling. See Appendix C for the full confusion matrix.

[11] Based on the data in Table 1 and repeated for visual comparison with the 2-point scale.

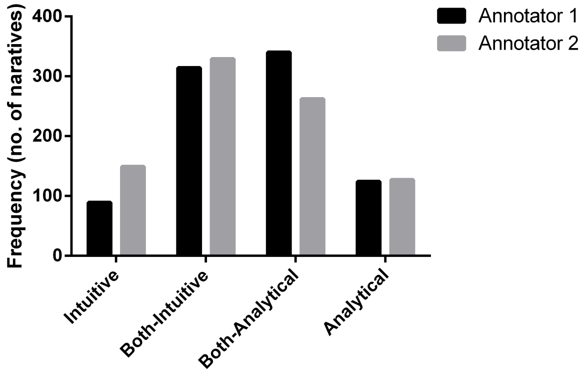


Figure 5. The distribution of ratings among the decision-making spectrum, on a 4-point scale.

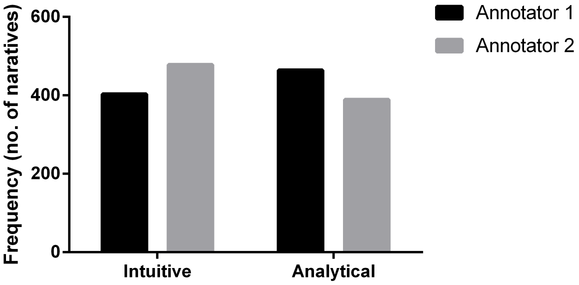


Figure 6. The distribution of ratings among the decision-making spectrum, on a 2-point scale.

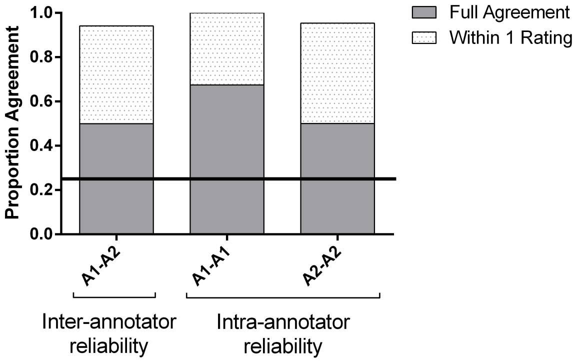


Figure 7. Inter- and intra-annotator reliability for the 4-point scheme, by proportion agreement. The reference line shows chance agreement (25%). *(A1=Annotator 1; A2=Annotator 2).*

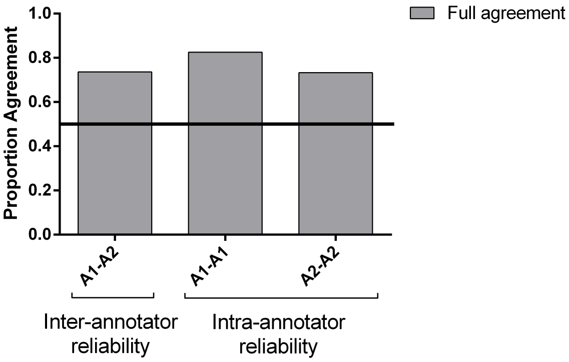


Figure 8. Inter- and intra-annotator reliability for the 2-point scheme, by proportion agreement. The reference line shows chance agreement (50%). *(A1=Annotator 1; A2=Annotator 2).*

the fuzziness of the categories and also that the subjective perception of decision-making style is systematic.

Annotator agreement was also assessed via linear weighted kappa scores (Cohen, 1968). As shown in Figure 9, inter-annotator reliability was moderate, and intra-annotator reliability was moderate (Annotator 2) to good (Annotator 1); see Landis and Koch (1977) and Altman (1991).

Since both proportion agreement and kappa scores were slightly higher for the 2-point scale, the automatic annotation modeling discussed below used this binary scale. In ad-

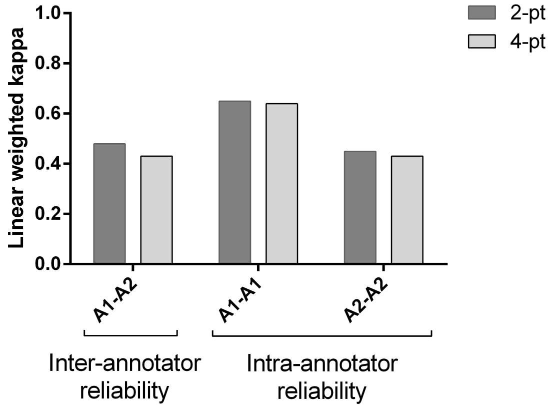


Figure 9. Annotator reliability, as measured by linear weighted kappa scores on the 2-pt and 4-pt scales.

dition, the distribution of data across binary classes was more balanced compared to the 4-point scale, as shown by the contrast between Figure 5 and Figure 6, further making it a suitable starting point for computational modeling.

Annotator strategy analysis. Five identical questionnaires (Appendix D) evenly spaced among the narratives asked annotators to rate how often they used various factors in judging decision style Table 2. Factors were chosen based on discussion with the annotators after the pilot, and referred to in descriptions of decision styles in the annotator instructions; the descriptions were based on characteristics of each style in the cognitive psychology literature (primarily based on two review papers: Evans, 2003 and Evans, 2008). Factors with high variability (*SD* columns in Table 2) reveal changes in annotator strategy over time, and factors that may influence intra-annotator reliability.

Both annotators reported using the *rel. (similarity) of final & first-mentioned diagnosis*, as well as *perceived attitude*, *perceived confidence*, and *use of justification*, to rate most narratives. Types of *processing* were used by both sometimes; this is important since these are central to the definitions of decision style in decision-making theory.

Differences in strategies allow for the assessment of annotators' individual preferences. Annotator 1 often considered the *no. of diagnoses in the differential*, and *rel. timing of the differential*, but Annotator 2 rarely attended to them; the opposite pattern occurred with

Table 2

Annotator Use of Factors in Rating Decision Style

	Annotator 1		Annotator 2	
Factor	*M*	*SD*	*M*	*SD*
Switching between decision styles	1.0	0.0	3.6	0.9
Timing of switch between decision styles	1.6	0.5	4.2	0.4
Silent pauses (...)	2.0	0.0	3.6	0.5
Filled pauses (e.g. *uh, um*)	2.0	0.7	3.6	0.5
Rel. (similarity) of final & differential diagnosis	2.8	0.4	3.2	0.8
Use of logical rules and inference	3.2	0.8	2.2	0.4
False starts (in speech)	3.4	0.9	2.4	0.9
Automatic vs. controlled processing	3.4	0.5	4.0	0.0
Holistic vs. sequential processing	3.6	0.5	4.4	0.5
No. of diagnoses in differential diagnoses	4.0	0.0	1.6	0.5
Word choice	4.0	0.7	2.6	0.5
Rel. (similarity) of final & first-mentioned diagnosis	4.0	0.0	4.0	0.0
Perceived attitude	4.0	0.7	4.0	0.0
Rel. timing of differential diagnosis in the narrative	4.2	0.8	2.8	0.8
Degree of associative (vs. linear, ordered) processing	4.2	0.4	3.8	0.4
Use of justification (e.g. *X because Y*)	4.2	0.4	4.0	0.0
Perceived confidence	4.4	0.5	4.2	0.4

Annotators rated each of the listed factors as to how often they were used in annotation, on a 5-point Likert scale from *for no narratives* (1) to *for all narratives (5).* This table shows the average and standard deviation over all 5 questionnaires, for each annotator. (Some factors slightly reworded.) *A1=Annotator 1, A2=Annotator 2.*

respect to *switching between decision styles*, and the *timing of the switch.*

The shared high factors reveal those consistently linked to interpreting decision style, despite the concept's fuzzy boundaries. In contrast, annotator differences in factor use reveal starting points for understanding fuzzy perception, and for further calibrating inter-annotator reliability.

Narrative case study. Examining particular narratives is also instructive. Of the 86 duplicated narratives with two ratings per annotator, *extreme agreement* occurred for 22 cases (26%), meaning that all four ratings were exactly the same (Figure 10). Notably, there were no cases where all four labels (primary + secondary, on the duplicated narratives) differed, and the distribution is similar to that of the primary rating distribution. Both of these points further emphasize the phenomenon's underlying regularity.

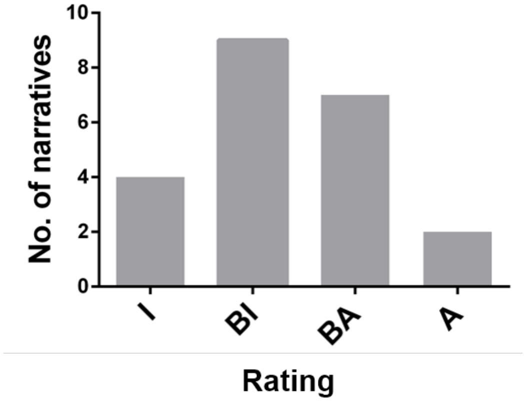
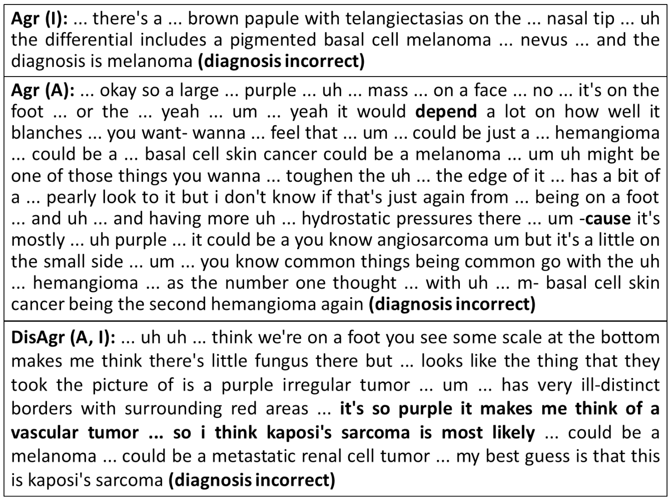


Figure 10. Distribution of duplicated narratives for which annotators exhibited extreme agreement on decision style rating. *I=Intuitive, BI=Both-Intuitive, BA=Both-Analytical, A=Analytical.*

Figure 11 (top) shows such a case of extreme agreement on intuitive reasoning: a quick decision without reflection or discussion of the differential. Figure 11 (middle) shows a case of analytical reasoning: consideration of alternatives and logical inference.



Agr (I): ... there's a ... brown papule with telangiectasias on the ... nasal tip ... uh the differential includes a pigmented basal cell melanoma ... nevus ... and the diagnosis is melanoma **(diagnosis incorrect)**

Agr (A): ... okay so a large ... purple ... uh ... mass ... on a face ... no ... it's on the foot ... or the ... yeah ... um ... yeah it would **depend** a lot on how well it blanches ... you want- wanna ... feel that ... um ... could be just a ... hemangioma ... could be a ... basal cell skin cancer could be a melanoma ... um uh might be one of those things you wanna ... toughen the uh ... the edge of it ... has a bit of a ... pearly look to it but i don't know if that's just again from ... being on a foot ... and uh ... and having more uh ... hydrostatic pressures there ... um -**cause** it's mostly ... uh purple ... it could be a you know angiosarcoma um but it's a little on the small side ... um ... you know common things being common go with the uh ... hemangioma ... as the number one thought ... with uh ... m- basal cell skin cancer being the second hemangioma again **(diagnosis incorrect)**

DisAgr (A, I): ... uh uh ... think we're on a foot you see some scale at the bottom makes me think there's little fungus there but ... looks like the thing that they took the picture of is a purple irregular tumor ... um ... has very ill-distinct borders with surrounding red areas ... **it's so purple it makes me think of a vascular tumor ... so i think kaposi's sarcoma is most likely** ... could be a melanoma ... could be a metastatic renal cell tumor ... my best guess is that this is kaposi's sarcoma **(diagnosis incorrect)**

Figure 11. Narratives for which annotators were in *full agreement* on *I* (top) and *A* (middle) ratings, vs. in *extreme disagreement* (bottom).

In the full dataset (initial ratings), there were 50 cases (6%) of 2-point inter-annotator disagreement and one case of 3-point inter-annotator disagreement (Figure 11, bottom). This latter narrative was produced by an attending (experienced physician), 40% confident and incorrect in the final diagnosis. Annotator 1 rated it analytical, while Annotator 2 rated it intuitive. This is in line with Annotator 1's preference for analytical ratings (Table 1). Annotator 1 may have viewed this pattern of *observation* → *conclusion* as logical reasoning, characteristic of analytical reasoning. Annotator 2 may instead have interpreted the phrase *it's so purple it makes me think of a vascular tumor...so i think [...]* as intuitive, due to the *makes me think* comment, indicating associative reasoning, characteristic of intuitive thinking. This inter-annotator contrast may reflect Annotator 1's greater reported use of the factor *logical rules and inference* (Table 2).

Automatic Annotation of Decision Style

This section describes the development of a computational model of decision style, with the goal of automatically annotating physician narratives. First, the process of data selection and labeling is described, based on the initial manual corpus annotation of decision style previously discussed. Next, the methods, features, feature selection, and results of automatic annotation are detailed. Finally, a study of feature combinations examines the relative contribution of various feature types towards decision style classification.

Data selection and labeling for computational modeling. This section details the systematic method used to select data for model development. One of the main goals of this work was to develop a computational model that could automatically annotate narratives as intuitive or analytical, based on lexical, speech, disfluency, physician demographic, cognitive, and diagnostic difficulty features. The study employed a supervised learning approach, and since no real ground truth was available, it relied on manual annotation of each narrative for decision style. However, annotators did not always agree on the labels, as discussed above. Thus, strategies were developed to label narratives, including in the case of disagreement (Figure 12).

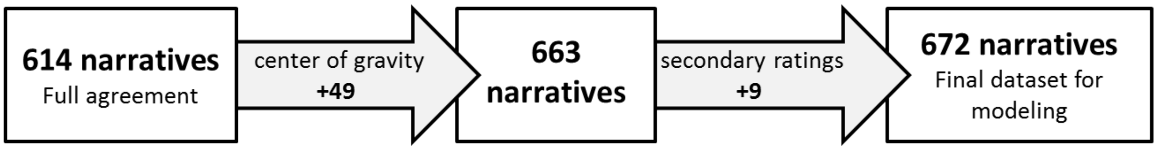
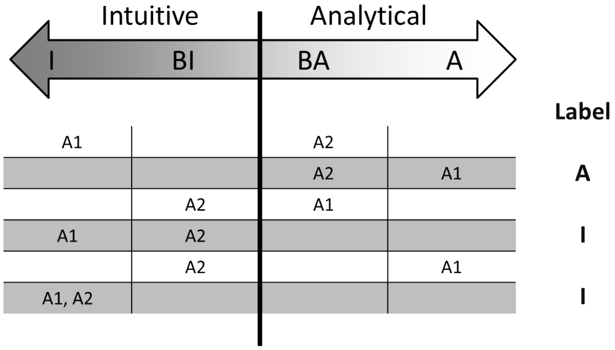


Figure 12. Narrative labeling pipeline. 614 narratives were labeled due to full binary agreement, and center-of-gravity and secondary rating strategies were used to label an additional 58 narratives for which annotators were not in agreement.

The dataset used for modeling consisted of 672 narratives. Annotators were in full agreement for 614 ratings on the binary scale of intuitive vs. analytical (Figure 13).[12] Next, 49 narratives were assigned a binary label based on the center of gravity of both annotators' primary ratings (Figure 14). For example, if a narrative was rated as *Intuitive* and *Both-Analytical* by Annotators 1 and 2, respectively, the center of gravity was at *Both-Intuitive*, resulting in an *Intuitive* label. Finally, 9 narratives were labeled using the annotators' secondary ratings,[13] available for 10% of narratives, to resolve annotator disagreement.[14]



I	BI	BA	A	Label
A1		A2		
		A2	A1	A
	A2	A1		
A1	A2			I
	A2		A1	
A1, A2				I

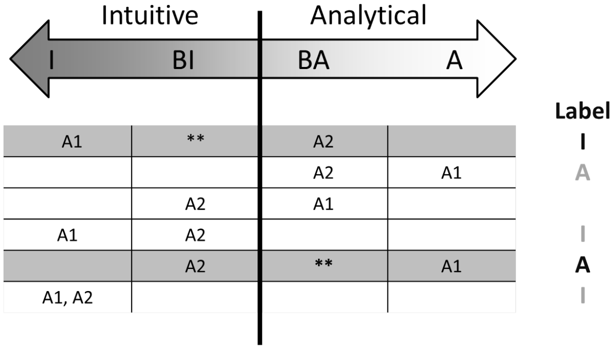
Figure 13. Demonstration of initial corpus labeling, in which 614 narratives were labeled on the basis of binary agreement.

Narratives with disagreements that could not be resolved in these ways were excluded. As perception of decision-making style is subject to variation in human judgment, this work focused on an initial modeling of data which represent the clearer-cut cases of decision style

[12] Excluding also narratives lacking confidence or correctness information.

[13] Collected to measure intra-annotator reliability.

[14] For example, if the primary ratings of Annotator 1 and Annotator 2 were *Both-Analytical* and *Both-Intuitive*, respectively, but both annotators' secondary ratings were intuitive (e.g., *Both-Intuitive* or *Intuitive*), the narrative was labeled *Intuitive*.



I	BI	BA	A	Label
A1	**	A2		I
		A2	A1	A
	A2	A1		
A1	A2			I
	A2	**	A1	A
A1, A2				I

Figure 14. Demonstration of center-of-gravity strategy, used to label an additional 49 narratives.

(rather than the disagreement gray zone on this gradient perception continuum). From the perspective of dealing with a subjective problem, this approach enables an approximation of ground truth, as a validation concept.

Methods. A model was developed for the binary prediction case (intuitive vs. analytical), since 2-point rating had slightly higher annotator agreement (see *Quantitative Annotation Analysis* above). Model development and analysis were performed using the WEKA data mining software package (Hall et al., 2009). The dataset was split into 80% development and 20% final test sets (Table 3).[15] Parameter tuning was performed using 10-fold cross-validation on the best features in the development set.[16]

Features. Three feature types were derived from the spoken narratives to study the linguistic link to decision-making style: lexical (37), speech (13), and disfluency (3) features. Three other feature types relevant to decision-making were demographic (2), cognitive (2), and difficulty (2) features (Table 4).

Relevant *lexical* features were extracted with the Linguistic Inquiry and Word Count (LIWC) software, which calculates the relative frequency of syntactic and semantic classes

[15] This split rests on the assumption that physicians may share common styles. Thus, the testing data will represent different physicians, but the styles themselves have been captured by the training data so that they can be correctly classified; the same rationale can be applied to image cases. To further investigate the phenomenon and identify the degree of inter- and intra-individual variation in decision style, future work could experiment with holding out particular images and physicians.

[16] In the feature combination study described below, parameters were tuned for each case of feature combinations in a similar way.

Table 3

Class Label Statistics

Label	80% Development Set	20% Final Test Set
Intuitive	276 (51%)	68 (51%)
Analytical	263 (49%)	65 (49%)
Total	539	133

in text samples based on validated, researched dictionaries (Tausczik & Pennebaker, 2010). *Disfluency* features were silent pauses, and the frequency of fillers and nonfluencies as computed by LIWC. *Speech* features are detailed in Table 4.

Besides linguistic features, three additional groups of features were included, with an eye towards application. *Demographic* features were gender and professional status, while *cognitive* features were physician confidence in diagnosis and correctness of the final diagnosis. *Difficulty* features consisted of an expert-assigned rank of diagnostic case difficulty, and the percent of correct diagnoses given by physicians for each image case, calculated on the development data only.

Feature selection. WEKA's CfsSubsetEval, an attribute evaluator, was used for feature selection,[17] using 10-fold cross-validation on the development set only. Features selected by the evaluator in at least 5 of 10 folds were considered best features. The best features from the entire feature set were: *2nd person pronouns, conjunctions, cognitive process, insight, cause, bio*, and *time* words, plus *silent pauses, speech length, time of min. pitch, standard deviation of pitch, time of min. intensity*, and *difficulty: percent correctness/image case.*

Feature selection, using the same attribute evaluator, was also performed on only the lexical features, which could be a starting point for analysis of decision-making style in text-only data. The best lexical features[18] included conjunctions, cause, cognitive process,

[17] With BestFirst search method.

[18] Best lexical features were: function words, singular pronouns, prepositions, conjunctions, quantifiers, and cognitive process, cause, discrepancy, tentative, inclusion, exclusion, perception, see, bio, motion, time, and assent words.

Table 4

Feature Types Used For Decision Style Modeling

Type	Feature	Description / *Examples*
Lexical	exclusion	*but, without*
	inclusion	*both, with*
	insight	*think, know*
	tentative	*maybe, perhaps*
	cause	*because, therefore*
	cognitive process	*know, whether*
	. . .	
Speech	speech length	number of tokens
	pitch	min, max, mean, st. dev., time of min/max
	intensity	min, max, mean, st. dev., time of min/max
Disfluency	silent pauses	number of
	fillers	*like, blah*
	nonfluencies	*uh, um*
Demographic	gender	male, female
	status	resident, attending
Cognitive	confidence	percentage
	correctness	binary
Difficulty	expert rating	ordinal ranking
	% correctness/image	percentage

The listed lexical features are a sub-sample of the total set. For a complete list of lexical features, see Appendix E.

inclusion, exclusion, and perception words. These lexical items seem associated with careful examination and reasoning, which might be more present in analytical decision-making and less present in intuitive decision-making. Some categories, especially inclusion (e.g., *with, and*), exclusion (e.g., *but, either, unless*), and cause words (e.g., *affect, cause, depend, therefore*), seem particularly good representatives of logical reasoning and justification, a key feature of analytical reasoning. But as shown in the next section, when available, speech and disfluency information is useful, and potentially more so than some lexical features.[19]

[19] Feature selection was also performed only on the linguistic (lexical, speech, and disfluency) features as a group. The best features of these types were: second personal pronouns, conjunctions, cognitive process, insight, cause, bio, and time words; silent pauses; and speech length, time of minimum pitch, standard

Decision style modeling results. Table 5 lists the results for the Random Forest (Breiman, 2001) and Logistic Regression (Cox, 1972) classifiers on the best features (as selected from all features) on the final test set, after training on the development set. Random Forest aggregates the results of multiple decision tree classifiers (Brieman, 2001), while Logistic Regression weights each feature to best explain the variance in a predicted binary variable (Cox, 1972). These results suggest that decision style can be quantified and classified on a binary scale; the percent error reduction (compared to baseline) for both classifiers is substantial.

Table 5

Decision Style Classifier Performance on Final Test Set

	Performance Metric			
Classifier	%Acc	%ER	Pr	Re
Random Forest	88	76	88	88
Logistic Regression	84	67	84	84
Majority Class Baseline	51	–	–	–

Performance on final test set; reduction in error (%ER) is calculated relative to majority class baseline. Precision (Pr) and recall (Re) are macro-averages of the two classes.

Feature combination exploration. A study of feature combinations was performed on the final test set with Random Forest (Table 6) to explore the contribution of each feature type towards automatic annotation. The best performance was achieved after applying feature selection on all features. Lexical and disfluency features were useful for determining decision style, and the best linguistic features (chosen with feature selection) were slightly more useful. These latter feature types improve on the performance achieved when considering only speech length and silent pauses, which were apparent characteristics to the human annotators and among the best features (see *Feature Selection* section).

Demographic features improved somewhat over the baseline, indicating an association

deviation of pitch, and time of minimum intensity. They could represent a starting for point for analyzing speech data not enhanced by additional speaker and task information.

Table 6

Feature Combination Study: Performance on Final Test Set

Features	Accuracy
All*	88
All	85
(Lexical + Speech + Disfluency)*	86
Lexical + Speech + Disfluency	84
Lexical + Disfluency	84
Only speech length and silent pauses	81
Disfluency	79
Lexical	77
Demographic + Cognitive	68
Demographic	64
Majority Class Baseline	51

Star (*) indicates the use of feature selection.

between gender, professional status, and decision-making, and adding cognitive features increased performance. Importantly, overall these findings hint at linguistic markers as key indicators of decision style.

Physician-Level Metrics of Decision Style and Metacognitive Awareness

Below, two metrics created for the assessment of physician decision style and metacognitive awareness are described. The decision style metric is based on annotator ratings for decision style, while the metric for metacognitive awareness is based on the relationship between confidence and accuracy for each physician. Both metrics are later used in the evaluation of the study hypotheses.

Physician profiles of decision style. Annotations were also used to characterize physicians' preferred decision style. A decision score was calculated for each physician as follows:

$$d_p = \frac{1}{2n}\sum_{i=1}^{n}(r_{A1_i} + r_{A2_i}) \tag{3}$$

where p is a physician, r is a rating, n is total images, and $A1, A2$ the annotators. Annotators' initial ratings were summed – from 1 for *Intuitive* to 4 for *Analytical* – for all image cases for each physician, and divided by 2 times the number of images, to normalize the score to a 4-point scale. Figure 15 shows the distribution of decision scores across *residents* and experienced *attendings*.

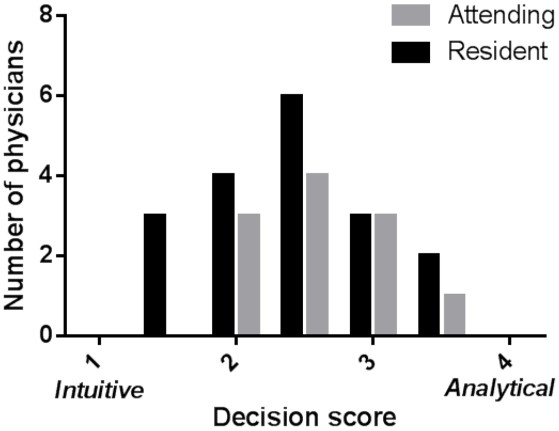


Figure 15. Distribution of physician decision scores by expertise.

Residents exhibited greater variability in decision style. While this might reflect that residents were the majority group, it suggests that differences in expertise are linked to decision styles; such differences hint at the potential benefits that could come from preparing clinical trainees to self-monitor their use of decision style. Interestingly, the overall distribution is skewed, with a slight preference for analytical decision-making, and especially so for attendings. Additionally, analyses of gender and decision style, diagnostic accuracy, and metacognitive awareness may be found in Appendix F.

Physician profiles of metacognitive awareness. Metacognitive awareness was computed for each physician (see Equation 2), based on the gamma correlation between their confidence estimates and diagnostic correctness across all image cases. Figure 16 shows the distribution of the gamma correlations among the 29 physicians. Gamma ranges from -1 to 1 (Bornstein & Zickafoose, 1999; Krug, 2007; Nelson, 1984), depending on the direction and strength of the confidence-accuracy relationship. As shown below, confidence and accuracy tended to be positively associated for the physician participants in this study, but the strength of this link varies among them.

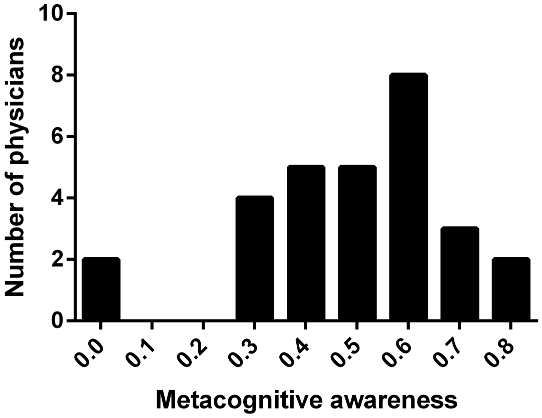


Figure 16. Distribution of metacognitive awareness among physicians, as measured by the gamma confidence-accuracy correlation.

The gamma correlation was chosen over the other relative measure of the confidence-accuracy relationship, the point-biserial correlation, since the gamma correlation is non-parametric, and thus makes less assumptions about the nature of the data (Krug, 2007). Second, the gamma correlation is considered the best measure of *resolution*, which captures individuals' tendency to have high vs. low confidence for accurate vs. inaccurate decisions at the level of the *individual* items or stimuli (Nelson, 1984). This is in contrast to other measures of the confidence-accuracy relationship, which consider individuals' overall level of confidence and compare it to their overall accuracy. Essentially, the gamma correlation captures the degree to which, for a particular item, an individual's confidence is predictive of accuracy.

Hypothesis Evaluation

Below, the results of hypothesis evaluation are reported (with the exception of *H1*, which corresponds to the decision style model discussed in the previous section). These hypotheses were evaluated using primary and secondary metrics based on raw annotation, confidence or accuracy data, including the decision score and metacognitive awareness metrics described in the previous section.

Table 7 lists and describes each metric, so as to serve as a reference for the following sections. Certain metrics were calculated based on the entire 867-narrative dataset, while others were calculated only using the 672-narrative dataset used for computational modeling.[20] This is because certain metrics take advantage of the binary (intuitive vs. analytical) labels applied to narratives in this latter set. Finally, the term *correct*, or alternatively *accurate*, as applied to physician narratives, is based on physicians' final diagnoses, as scored by a licensed dermatologist.[21]

Decision style. Three hypotheses concerned physician decision style, with respect to the relationship between decision style and diagnostic accuracy, expertise, and case difficulty. Since there are multiple possible metrics for decision style (see Table 6), as well as multiple levels of analysis (physician vs. narrative), some hypotheses are evaluated with several analyses.

H2. *Intuitive reasoning will be associated with lower levels of accuracy.*

This hypothesis was evaluated from two perspectives. The narrative level considered the relationship between the use of intuitive reasoning and diagnostic accuracy with each physician narrative as the basic unit of analysis, across physicians and cases. In contrast, the physician level considered the relationship between intuitive reasoning and accuracy for each physician, to answer the question: do physicians who tend towards intuitive reasoning have lower rates of diagnostic accuracy?

Narrative-level analysis. First, at the narrative level, a frequency table was created, comparing decision style and accuracy across physicians and image cases. Table 8 shows the relationship of physicians' diagnostic correctness by decision style (intuitive vs. analytical on a binary scale), given the 672-narrative dataset.

[20] See the section titled *Data Selection and Labeling for Computational Modeling* for more detail.

[21] See the section titled *Preliminary Data Analysis* for more information.

Table 7

Summary of Case and Physician Metrics

Computed For	Metric	Description	Data
Physician	metacognitive awareness	gamma confidence-accuracy correlation across all cases for each physician (see Eq. 2)	867
Physician	physician decision score	averaged, normalized metric based on raw decision ratings for each physician (see Eq. 3)	867
Physician	physician proportion correct	proportion of correct narratives for each physician	867
Physician	physician analytical proportion	proportion of analytical narratives for each physician	672
Physician	intuitive-correct proportion	proportion of correct narratives, out of total number of intuitive narratives, for each physician	672
Physician	analytical-correct proportion	proportion of correct narratives, out of total number of analytical narratives, for each physician	672
Case	case difficulty	percent correct for each image case, across all physicians (see Eq. 4)	867
Case	case decision score	averaged, normalized metric based on raw decision ratings for each case	867
Case	case analytical proportion	proportion of analytical narratives for each case (based on binary narrative labels)	672

In the *Data* column, *672* indicates the dataset used for computational modeling, with use of binary decision labels where appropriate; *867* indicates use of the entire set of narratives, with use of 4-point decision ratings where appropriate.

Overall, there was a slightly higher prevalence of intuitive reasoning, and there were more incorrect than correct diagnoses.[22] Table 8 also suggests a relationship between correctness and decision-making style, where for correct diagnoses, intuitive reasoning was more dominant. The opposite trend held for incorrect diagnoses: analytical reasoning was more frequent. Indeed, a chi-square test revealed a significant relationship between correctness and decision style, $\chi^2(1,\, N = 672) = 13.05$, $p < 0.01$.

[22] Contributing factors to the proportion of incorrect diagnoses might include case difficulty levels in the experimental scenario, and that physicians did not have access to additional information, such as patient history or follow-up tests.

Table 8

Distribution of Diagnostic Correctness by Decision Style

	Correctness		
Decision style	Correct	Incorrect	Total
Intuitive	158	186	344
Analytical	106	222	328
Total	264	408	672

Physician-Level Analysis. Two physician-level analyses were performed. The first was based on the binary decision labels, while the second was based on the decision score computed for each physician based on annotator ratings on the 4-point rating scale.[23]

For the first analysis, two new metrics were created for each physician, also based on the 672-narrative dataset. The first, *intuitive-correct proportion*, was the proportion of correct narratives, out of the total number of intuitive narratives. The second, *analytical-correct proportion* was the proportion of correct narratives, out of the total number of analytical narratives. Table 9 illustrates the correctness by decision style categorization. The intuitive-correct proportion was calculated as a/c, and the analytical-correct proportion was calculated as d/f.

Table 9

Schematic Representation of Decision Style by Correctness

	Correct	Incorrect	Total
Intuitive	a	b	c
Analytical	d	e	f

An intuitive-correct proportion (a/c) and analytical-correct proportion (d/f) were computed for each physician.

[23] For more information, see the section titled *Physician Profiles of Decision Style.*

Both the the intuitive-correct proportion and analytical-correct proportion were computed for each physician. Next, before any statistical analysis, a frequency histogram was constructed to examine the trend (Figure 17). As apparent in the figure, intuitive reasoning was clearly associated with higher accuracy, opposite to the anticipated trend, so no further hypothesis testing was necessary. However, a two-tailed Mann-Whitney test showed a significant difference between the intuitive-correct (Mdn = 0.50) and analytical-correct distributions (Mdn = 0.29), U= 172, $p < 0.001$. This shows a significant trend in the opposite (than anticipated) direction, linking intuitive decision-making to greater diagnostic accuracy.

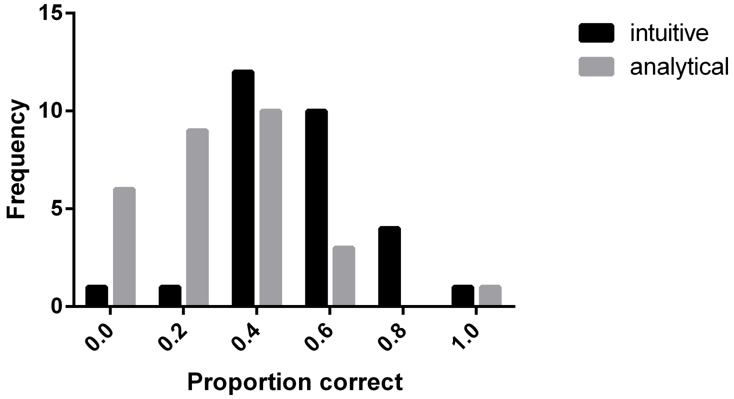


Figure 17. Distribution of correctness at the physician level, comparing the intuitive-correct proportion and analytical-correct proportion.

A second analysis used the decision score calculated for each physician, and compared it to each physician's overall accuracy (number of image cases correct/total number of image cases). These metrics were used to evaluate, in another way, whether physicians who tend towards intuitive reasoning are generally more accurate in their diagnosis. This evaluation, due to the nature of the metrics, utilized all 867 data points.

Figure 18 shows the results of this analysis. There is no apparent correlation between decision score and physician correctness. Indeed, a Spearman correlation between the two variables was not different from zero, $r_s(29) = .13$, $p = .516$.

Thus, the narrative-level analysis and the first physician-level analysis suggested that

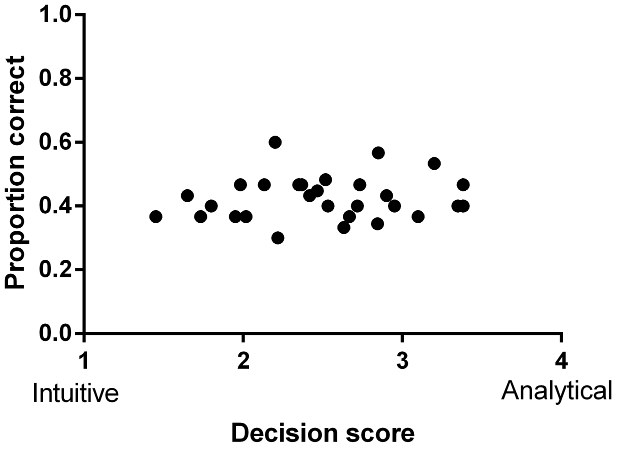


Figure 18. Relationship between physician decision score and physician proportion correct.

intuitive reasoning was linked to higher accuracy, rather than lower accuracy as anticipated. The second physician-level analysis showed no link between decision style and diagnostic accuracy. Taken together, all three analysis failed to confirm Hypothesis 2.

H3. *Experts will have better success with intuitive reasoning than novices.*

To examine this hypothesis, a between-group comparison was performed for residents vs. attendings (experts). Success with intuitive reasoning was defined for each physician as the *intuitive-correct proportion*: the proportion of cases that were diagnosed correctly using intuitive reasoning, out of all of the cases that were diagnosed using intuitive reasoning.[24] Intuitive reasoning was defined based on the binary labels used in computational modeling, which restricted this analysis to the 672-narrative dataset.

Figure 19 shows the distribution of the residents vs. attendings (experts), with respect to the intuitive-correct proportion metric. The graph shows that attendings (Mdn = 0.55) had better success with intuitive reasoning than residents (Mdn = 0.44), and a Mann-Whitney test revealed a significant difference between the distributions, U= 42, p = .009. Thus, this result supports Hypothesis 3.

H4. *More difficult cases will be associated with analytical decision-making, while less*

[24] This metric was also used to evaluate Hypothesis 2; see Table 9.

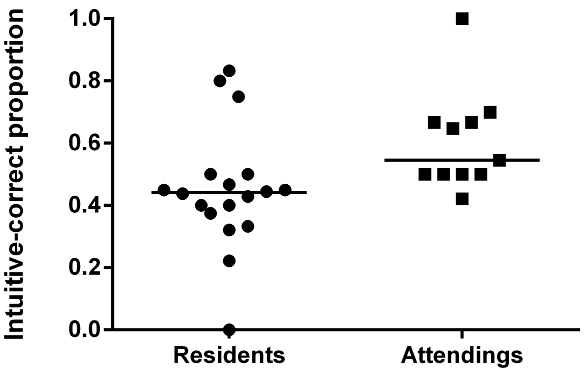


Figure 19. Success in using intuitive reasoning, by expertise. The horizontal line indicates the median of each distribution.

difficult cases will be associated with intuitive decision-making.

To examine this hypothesis, the relationship between case difficulty and decision style was evaluated. *Case difficulty* was measured by the percent of correct diagnoses for each image case, over all physicians.[25] Thus, this metric was highest (100%) for easy image cases, which all physicians diagnosed correctly, and lowest (0%) for difficult image cases, which no physicians diagnosed correctly.

Decision style was measured in two ways, so two analyses were performed, one for each decision style metric. First, a decision score was created for each case, similar to the physician decision score. The *case decision score* was calculated as follows:

$$d_c = \frac{1}{2n}\sum_{i=1}^{n}(r_{A1_i} + r_{A2_i}) \tag{4}$$

where c is a case, r is a rating, n is total images, and $A1, A2$ the annotators. Annotators' initial ratings were summed – from 1 for *Intuitive* to 4 for *Analytical* – for all physicians for each image case, and divided by 2 times the number of cases, to normalize the score to a 4-point scale. Case decision score was calculated based on the entire dataset (867 annotated narratives). Since it is a continuous score based on raw annotator ratings, is likely be a better estimate of case decision score than a metric based on the binary labels used for

[25] This metric was also used in the computational model of decision style. The other available difficulty metric, the expert dermatologist rating of difficulty, was not used to evaluate this hypothesis, as it is an ordinal rating and is likely to be more subjective than a performance-based metric based on many individuals.

computational modeling of decision-making.

Figure 20 shows the relationship between case difficulty and case decision score. As case difficulty increases, the case decision score tends towards analytical reasoning. A Spearman correlation between the two measures was significant, $r_s(30)$ = -.51, p = .004.

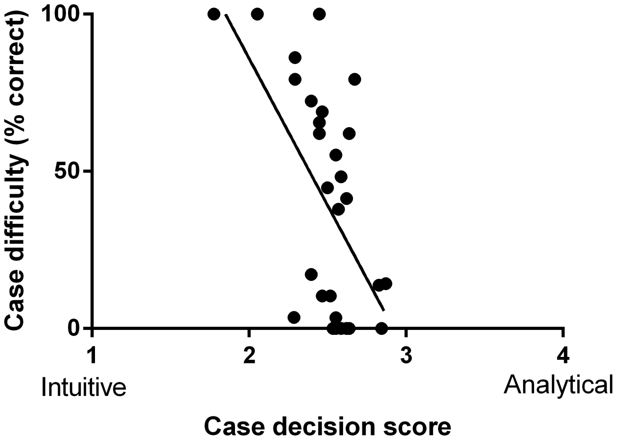


Figure 20. Relationship between case decision score and case difficulty.

The second measure of decision style was the *case analytical proportion*, defined as the proportion of analytical ratings for each case across all physicians. As case analytical proportion is based on the binary decision labels, it was computed based on the 672-narrative dataset used for computational modeling. Figure 21 shows the relationship between case analytical proportion and case difficulty. As the case difficulty increases, the use of analytical reasoning does as well. Indeed, a Spearman correlation between the two measures was significant, $r_s(30)$ = -.54, p = .002. Thus, both analyses of the link between case difficulty and decision style support Hypothesis 4.

Metacognitive awareness. The two hypotheses on metacognitive awareness concern the link between it and diagnostic accuracy, as well as expertise. For both hypotheses, metacognitive awareness was computed for each physician as the gamma correlation between confidence and accuracy, across all cases diagnosed by the physician. For more information on gamma, see the section *Physician Profiles of Metacognitive Awareness.*

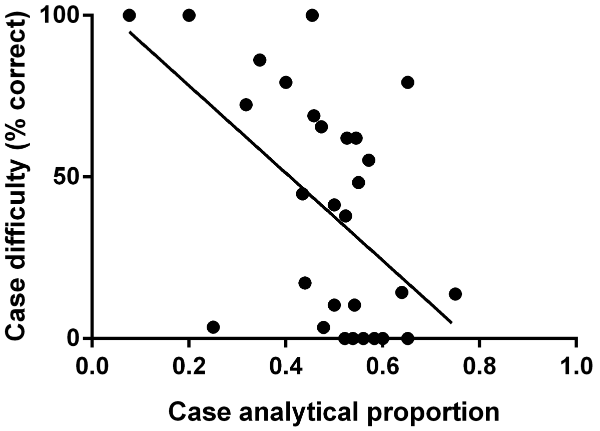


Figure 21. Relationship between case difficulty and case analytical proportion.

H5. *Higher levels of metacognitive awareness will be associated with higher levels of diagnostic accuracy.*

This hypothesis was evaluated at the physician level, since only individuals (rather than cases) can possess metacognitive awareness. To evaluate this hypothesis, *metacognitive awareness* was defined as the confidence-accuracy correlation for each physician, as discussed above. Diagnostic accuracy was defined for each physician as the *physician proportion correct*: the proportion of cases that the physician diagnosed correctly. Since neither metric relies on binary decision labels, the larger dataset of 867 narratives was used to compute each metric.

Figure 22 shows the relationship between metacognitive awareness and physician proportion correct. The two measures were negatively correlated, $r_s(29)$ = -.40, p = .032. This trend was in the opposite direction than anticipated, so this result fails to confirm Hypothesis 5.

H6. *Experienced physicians will exhibit higher levels of metacognitive awareness than inexperienced physicians.*

This hypothesis was assessed by comparing metacognitive awareness, as defined by the confidence-accuracy gamma correlation, in the resident vs. attending (expert) group. Fig-

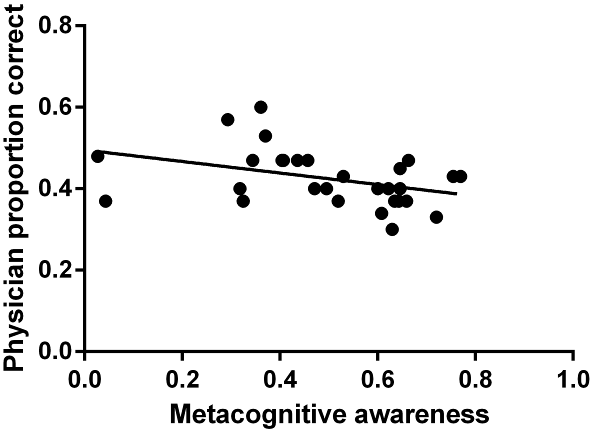


Figure 22. Relationship between metacognitive awareness and physician proportion correct.

ure 23 shows the distribution of metacognitive awareness among residents (*Mdn* = .525) and attendings (*Mdn* = .495). A Mann-Whitney test revealed no significant difference between the two distributions, U= 86.5, p = .588. Thus, this result fails to confirm Hypothesis 6.

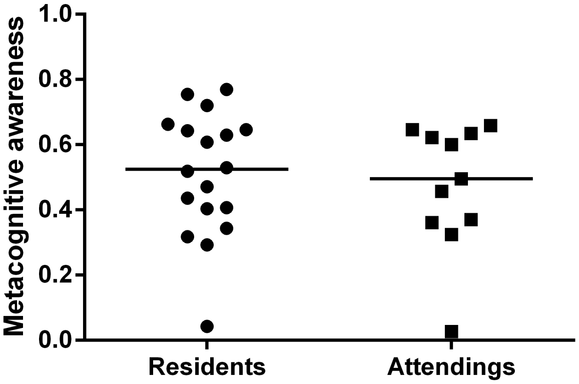


Figure 23. Metacognitive awareness by expertise. The horizontal line indicates the median for each group.

Decision style and metacognitive awareness. This final hypothesis considers the link between decision style and metacognitive awareness.

H7. *Higher levels of metacognitive awareness will be associated with increased use of analytical decision-making.*

This hypothesis was evaluated by examining the relationship between metacognitive awareness and decision style. As previously, metacognitive awareness was defined by each physician's gamma confidence-accuracy correlation, while decision style was computed according to two different metrics.

The first decision style measure used was physician decision score, which provides a general assessment of a particular individual's decision style across all cases s/he diagnosed.[26] Figure 24 shows the relationship between metacognitive awareness and physician decision style. The two measures were not significantly correlated, $r_s(29) = .18$, $p = .360$.

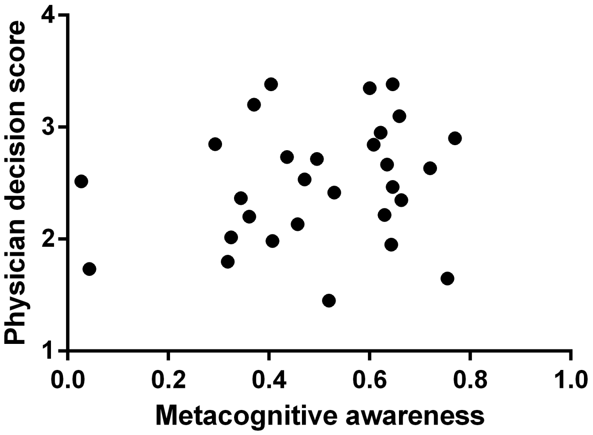


Figure 24. Relationship between metacognitive awareness and physician decision score.

The second decision style measure used was the *physician analytical proportion*, calculated for each physician based on the 672-narrative dataset used for computational modeling. For each physician, the proportion of narratives labeled *analytical* was computed, out of all the physician's narratives.[27] Figure 25 shows the relationship between metacognitive awareness and physician analytical proportion. The two measures were not significantly correlated, $r_s(29) = .17$, $p = .366$. Together, both analyses fail to support Hypothesis 7.

[26] See the section titled *Physician Profiles of Decision Score* for more information on this metric.

[27] Equally, the corresponding *physician intuitive proportion* could have been computed and used for this analysis, as the decision labels are binary, so that both proportions together make up the entire whole.

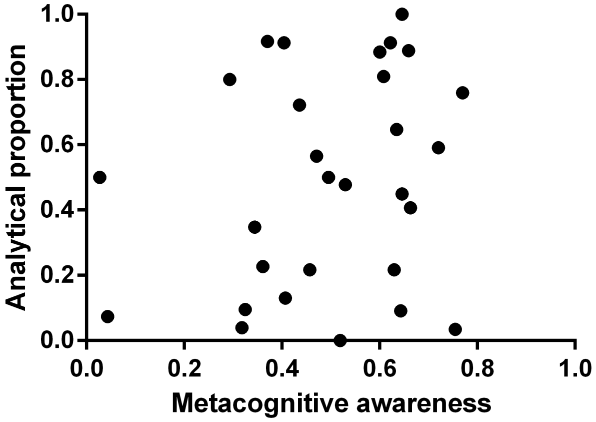


Figure 25. Relationship between metacognitive awareness and physician analytical proportion.

Discussion

Decision Style Annotation and Modeling

In this work, annotators showed a systematic perception and moderately reliable annotation of decision style, which was successfully detected for each narrative at substantial improvement over the baseline. Thus, Hypothesis 1 – *Decision style can be reliably annotated in from physician verbalizations in clinical reasoning contexts, and used to create a computational model for the automatic annotation of such verbalizations* – was supported. The fact that decision style could be annotated reliability, and that such annotation could be used, further, for modeling and automated annotation prediction, lends validity to the dual process theory (Evans, 2003; Kahneman & Frederick, 2002; Stanovich & West, 2000). It also lends support to claims that cognitive processes, particularly decision style, are revealed in language use (Pennebaker & King, 1999; Tausczik & Pennebaker, 2010).

The study leaves exploration of finer-grained computational modeling at the 4-point scale for future work. Further study could also focus on determining the optimal rating scale for decision annotation, which approaches the limits of annotators' ability to reliably detect reasoning style. In this work, the rating scale was a 4-point rather than 5-point rating scale, so as to force annotators to decide whether intuitive or analytical reasoning was more prominent. However, it may be the case that a 5-point, 6-point, or even 7-point scale of decision style may be appropriate in some contexts. Such a scale would provide higher resolution for statistical analyses linking decision style to performance. Alternatively, annotators may be asked to suggest their own annotation scale, as in Burstein and Chodorow (2014), in which essays were annotated for coherence, also a subjective task. Aggregation and analysis of such scales over multiple annotators may give additional insight into mental models of decision style as well as the limits of resolution with respect to decision style judgment.

Based on the results of corpus annotation, it is clear that annotators consider a range of factors in decision style annotation. To improve future corpus annotation, an iterative annotation process could be used, in which annotators work on smaller portions of the corpus, in stages of annotation-evaluation-adjustment, with re-annotation based on additional training, or even adjustments to the rating scale, when necessary (Pustejovsky and Stubbs, 2012). In this work, the pilot and the main annotation represent two annotation cycles, which is preferable to only one cycle; and in fact, the discussions with annotators yielded the basis for the annotator questionnaire and several other key adjustments to corpus annotation. Finally, another, more expensive way of ensuring good corpus annotation is to use more than two annotators. Such a study might even compare the inter-annotator reliability between experts in cognitive psychology, medicine, and naive annotators without expertise in either domain, so as to shed light on the relative importance of domain expertise in clinical decision annotation, and the extent to which more expensive expert annotation is necessary.

Based on the feature combination study, linguistic features were more useful in prediction of binary decision style than demographic and cognitive features. This finding supports the use of post-hoc analyses of existing language data for decision style, via both manual and automatic annotation of decision style, so as to analyze the relationship between decision style and accuracy in particular contexts and domains, even when demographic and cognitive information is not available. Disfluency features, particularly silent pauses, were also important in decision style prediction. These results also align with Womack et al. (2012), who proposed that silent pauses in physician narration may indicate cognitive processing. Thus, this pattern of results may be due to the fact that analytical decision-making may recruit more cognitive resources than intuitive decision-making. There is also an interesting parallel to other fairly recent findings that subjective disfluency cues (in the non-linguistic sense) prompt the use of analytical reasoning (Alter, Oppenheimer, Epley, & Eyre, 2007). Future work might investigate the link between subjective experiences of disfluency and disfluencies in language, as, taken together, these studies may suggest that the two may be positively correlated. That is: individuals who experience feelings of disfluency or difficulty may be more likely to utter disfluent tokens, so that both may be positively correlated with use of analytical reasoning in decision-making contexts. If this prediction is borne out, then disfluent tokens in language may be a useful proxy for subjective experiences of difficulty and disfluency, with the advantage that naturally elicited speech is an online, non-invasive form of data that avoids biases resulting from directly querying participants.

Finally, the distribution of decision annotation ratings showed that most clinical decision-making occurs in the central part of the continuum. This result is in line with the Lauri et al. (2001) study, in which nurses in five countries were asked to rate statements representative of intuitive or analytical decision-making on a 5-point scale. The authors found that reasoning varies with context and that styles in the middle of the cognitive continuum predominate. This result may be interpreted in a positive light, since it shows that clinicians exhibit flexibility in the process of decision-making. In fact, it has been suggested that clinicians may benefit from recruiting both System 1 and System 2 for the same diagnostic case, and thus reap the benefits of each (Norman, 2009).

Decision Style and Metacognitive Awareness

Decision style and diagnostic accuracy. Hypothesis 2 – *Intuitive reasoning will be associated with lower levels of diagnostic accuracy* - was not supported. In fact, the opposite trend was observed: intuitive reasoning was linked to greater levels of diagnostic accuracy. This result, while not in the anticipated direction, sheds light on the debate over the accuracy of System 1 vs. System 2. Although System 1 has been linked to the inappropriate use of heuristics and biases, which may decrease diagnostic accuracy (Croskerry, 2003b), it has also been linked to claims that intuitive reasoning is linked to better performance when much information is to be processed. In this view, mechanisms of intuitive reasoning and pattern recognition allow individuals to overcome the limitations of their working memory (Evans, 2008).

Viewed from the perspective of cognitive continuum theory, the higher prevalence of incorrect diagnoses for physicians using analytical decision style may be due to a mismatch between reasoning style and the task demands of the particular case (Hammond, 1981).

Finally, it might be the case that diagnostic difficulty was a moderating variable, where physicians preferred intuitive reasoning for less challenging cases, and analytical reasoning for more difficult cases. In fact, this connection between difficulty and decision style was observed with respect to Hypothesis 4, discussed below.

It is notable that the binary decision-based analysis supported the link between intuitive decision style and increased diagnostic accuracy, while the physician decision score was not linked to diagnostic accuracy in either direction. This may be because physician decision scores were computed as a normalized average across all cases, so that most scores were clustered in the middle range of the decision continuum and, thus, were not predictive due to low variance. This difference may also be related to the fact that the analyses were done on different-sized datasets, which may have varied systematically (e.g., in which certain cases were represented more than others, and/or in the case that narratives for which annotators disagreed shared some key characteristics).

Intuitive reasoning, diagnostic accuracy, and expertise. Hypothesis 3 – *Experts will have better success with intuitive reasoning than novices* – was confirmed. This is based on the significant difference between the distributions of correct diagnoses when using intuitive reasoning, among residents vs. attendings (expert physicians). This observation is line with the fact that experts have a broader base of experience to draw on, and is also in line with Klein's theory of recognition-primed decision-making, by which experts are able to quickly assess a situation and determine a response based on pattern recognition mechanisms developed over years of experience (Klein, 1999). These results are also in line with a recent empirical study, which found that expert basketball players had more success using intuitive reasoning than non-experts when asked to judge the difficulty of basketball shots (Dane, Rockmann & Pratt, 2012). In the same study, experts were also better than non-experts at recognizing authentic vs. counterfeit handbags when using intuitive reasoning (Dane et al., 2012). Taken together, these studies and the current work are notable in that all three experimental tasks incorporate perceptual expertise, and all three found that, with respect to intuitive decision-making, experts have more success than non-experts. In an applied setting, these results may suggest it is beneficial to encourage experts to trust their intuition, at least under certain conditions. This may be especially relevant in domains where analytical reasoning is valued or where explicit, post-hoc rationalizations of mission-critical decisions are necessary in investigative reports, both of which may encourage the use of analytical reasoning even when it is sub-optimal.

Case difficulty and decision style. Hypothesis 4 – *More difficult cases will be associated with analytical decision-making, while less difficult cases will be associated with intuitive decision-making* – was confirmed, based on two analyses, each of which used a different decision metric (case decision score and case analytical proportion). These results are in line with claims by Alter et al. (2007) that experiences of disfluency and difficulty are linked to greater use of analytical decision-making, assuming that more difficult cases did in fact prompt such experiences in the physician participants in this study. In addition, this link between difficulty and decision style may explain the results of Hypothesis 3, in which intuitive reasoning was linked to greater diagnostic accuracy. If physicians tend to use intuitive reasoning for easier cases, it is no surprise that they are also more likely to be correct in those instances. Since the original experimental study did not systematically manipulate case difficulty – rather, case difficulty was a measure derived from physician performance – the extent to which difficulty moderates the link between reasoning style and diagnostic accuracy cannot be determined. However, future studies on clinical decision-making might systematically manipulate both decision style and difficulty in order to determine the nature of the interactions between decision style, diagnostic difficulty, and diagnostic accuracy.

Metacognitive awareness and diagnostic accuracy. Hypothesis 5 – *Higher levels of metacognitive awareness will be associated with higher levels of diagnostic accuracy* – was not confirmed. In fact, a significant trend was observed in the opposite direction, in which higher levels of metacognitive awareness were linked to lower levels of diagnostic accuracy. This result was surprising, and contrasts with previous assertions that metacognitive awareness is linked to increased performance in the domains of reading and mathematics (e.g., Paris & Oka, 1986; Tobias & Everson, 1995). Also in clinical settings, there is some preliminary evidence, from laboratory settings, that metacognitive interventions based on conscious reflection have the potential to help individuals correct initial misdiagnosis (Coderre, Wright, & McLaughlin, 2010) and somewhat reduce availability bias (Mamede, van Gog, & van den Berge K, 2010). However, the vast majority of proposed cognitive interventions to reduce diagnostic error, a category which includes metacognition-based intervention, have either never been tested, or not been tested out of the laboratory (Graber et al., 2012).

This finding on the link between greater metacognitive awareness and decreased diagnostic accuracy should be taken with some reservation, since it contrasts to claims from both empirical and theoretical literature that metacognition is positively correlated with performance. It may be the case that, in this study, the measure of metacognitive awareness, as it was based on the confidence-accuracy relationship for each physician, suffered from sub-optimal estimates of its two components. With respect to confidence, physicians used the scale inconsistently (see *Limitations* section, below) and may have varied with respect to internal notions of low vs. high confidence. In addition, the relatively high prevalence of incorrect diagnoses, higher than misdiagnosis rates in general clinical practice,[28] also lead to inaccurate estimates of diagnostic accuracy. Together, these effects may have skewed the measure of metacognitive awareness in this study, so that it was not an accurate estimate of physicians' true metacognitive awareness.

Another factor with potential impact on physicians' metacognitive awareness is the nature of the experimental task. Since in clinical contexts, dermatologists generally have access to additional information, such as patient history, and can also request additional information from the patient and order follow-up tests, they are likely to calibrate their metacognitive awareness regarding diagnosis with respect to the actual clinical environment. Accordingly, it is possible that that the physician participants in this study, since they had less experience and feedback on their performance in diagnosing based solely on images, had skewed or lower confidence-accuracy calibration than they may have in professional contexts. Future work might investigate the links between metacognitive awareness and the

[28] Estimated at 5-15%, depending on the specialty (Berner & Graber, 2008).

amount and type of information available to physicians, and the ecological validity of the experimental context.

Metacognitive awareness, diagnostic accuracy, and expertise. Hypothesis 6 – *Experienced physicians will exhibit higher levels of metacognitive awareness than inexperienced physicians* – was not confirmed, based on the similar distributions of metacognitive awareness in the resident vs. attending groups. This is somewhat surprising, as metacognition has been deemed to key to the development of expertise, both generally (Sternberg, 1998) and in medicine (Quirk, 2006). In addition, improvements in metacognitive awareness have been linked to immediate feedback (El Saadawi et al., 2010), to which experts have likely had greater exposure over a lifetime of practice. On the other hand, it may be the case that, once attending physicians leave residency, they no longer get immediate feedback about clinical decisions from expert physicians, so there is no related increase in metacognitive awareness. Also, the El Saadawi et al. study was performed in the context of an instructional system, so such findings may not translate to clinical contexts in the long term.

In addition, it may be that metacognitive awareness is a relatively stable trait, related more to an individuals' propensity towards analytical thought (Thompson, 2009), rather than expertise. Also, as discussed with respect to Hypothesis 5, this finding may be due to biases affecting the measurement of metacognitive awareness via the confidence-accuracy relationship. Finally, this failure to support the hypothesis may be due to the small available sample size, as each group of physicians consisted of less than 20 individuals. Perhaps the effect, if it exists, can only be detected in large sample sizes; this is an additional avenue for future research.

Metacognitive awareness and decision style. Hypothesis 7 – *Higher levels of metacognitive awareness will be associated with increased use of analytical decision-making* – was not confirmed, based on analyses linking metacognitive awareness to two different metrics of decision style (physician decision score, averaged over all cases for each physician, and use of physician use of analytical decision-making, based on binary decision labels). As discussed above, this failure to confirm the hypothesis is consistent with several explanations: either that the metrics used for decision style and metacognitive awareness were inaccurate estimates of each; that sample size was insufficient to detect and effect; or that there is in fact no effect. With respect to this latter scenario, if Alter et al. (2007) are correct that difficulty (and other contextual cues) prompt the use of analytical decision-making, it may be the case that such cues are so salient that an individuals' level of metacognitive awareness does not mediate this difficulty-decision style link. However, even if this explanation is borne out empirically, metacognitive awareness may also serve decision-making through other processes and mechanisms (see Thompson, 2009, for a review). Thus, future research should systematically manipulate and measure decision style and metacognitive awareness so as to better determine the nature of the relationship between them, particularly with respect to clinical reasoning.

Limitations

Using a secondary dataset, while not uncommon in academic research, does have certain limitations, as the study was not originally designed to answer the research questions posed in the current study. The first concern is with respect to ecological validity: in real-world medical contexts, physicians diagnose not only on the basis of visual information but on the basis of lab results, vitals, patient demographics, patient risk factors and lifestyle, and other information. Based on their differential diagnosis, physicians can order tests and follow-up visits before determining a final diagnosis. In the study task, however, physicians made diagnoses on the basis of limited visual information, so that their diagnoses may represent only part of the overall clinical decision-making process. This may be reflected in the relatively high rate of incorrect diagnoses, though that may also be due to the inclusion of residents, who are still in training, in the study. In addition, the master-apprentice scenario asks physicians to describe each image case as if teaching a student, so the narratives may reflect teaching processes as well as decision-making process. Finally, since participant physicians also varied with respect to their professional and educational backgrounds, these training differences might have been another unmeasured source of variance affecting decision style. Future work might study the effects of training lineages on clinical decision style systematically, or perhaps attempt to create a relational hierarchy classifying and categorizing the various approaches in relation to theoretical frameworks of decision-making.

This work also relies on the assumptions that verbal data reflect working memory (Ericsson & Simon, 1993), and that cognitive processes are revealed in language use (Pennebaker & King, 1999; Cohn, Mel & Pennebaker, 2004). However, diagnosis, particularly in a visual medical specialty such as dermatology, also relies on visual attention and perceptual processes (Anderson & Shyu, 2011). Thus, it is reasonable to assume that these processes are not necessarily reflected in physicians' spoken narratives.

Another concern is with respect to the measurement of intuitive vs. analytical decision-making. It is possible that while there is some variability in participants' decision-making, their most intuitive and most analytical exemplars may not in fact be representative of the far ends of the spectrum. It is possible, then, that the reasoning reflected in the current study spans the middle of the intuitive-analytical spectrum, so that reasoning considered intuitive or analytical in the current experiment is only partially reflective of reasoning that is purely intuitive or purely analytical. However, since the variability within clinician reasoning in the current dataset could be measured reliably by both human and computational classification, then at least some features of intuitive vs. analytical reasoning were present in the narratives. Finally, the LIWC software used for lexical features considers surface strings rather than their conceptual senses; future work might operate on the sense rather than token level, and may also consider discourse structure in the narratives.

In addition, certain features of the task used in the current study may have induced particular types of reasoning. In his discussion of cognitive continuum theory as applied to clinical reasoning, Hamm (1988) notes that specific tasks which are "presented in a manner that guide the doctor to address a sequence of subtasks...will induce analytical cognition" (p.6). Thus, the fact that the task required participants to provide a case description, differential diagnosis, and then a final diagnosis may have influenced participants to use an analytical decision-making process. However, Hamm also notes that "if the information is presented pictorially, it induces intuition" (p.6). Thus, the images provided as stimuli in the study may have induced intuitive reasoning among participants. Finally, Hamm notes that explicit or implicit time pressure may induce intuitive thinking: "if only a brief time is available, the doctor will adopt intuitive cognition" (p.6). This may have implications for the current study as well. While the task was not performed under any explicit time pressure, and physicians cued when they were ready the next image case, they may have still been under some implicit time pressure. This is because the participants were generally informed of the total task duration of around 30-45 minutes, and were also told in the experimental instructions that the task would include 30 images. Thus, participants may have deduced that they had about a minute per image case, and performed the task under this unstated expectation. In addition, participants may have been eager to complete the task quickly and return to other responsibilities.

Finally, another concern is with respect to confidence scores, which were used to estimate metacognitive awareness. Clinicians may even be trained, implicitly or explicitly, not to show uncertainty, resulting in an inflation of confidence scores reported (see Katz, 1984, for a review). In fact, Croskerry and Norman (2008) note that overconfidence is the most significant bias in clinical decision-making. In this study, in fact, physicians tended to use only the upper range of the confidence scale, in line with these claims.

Applications

The decision annotation scale and computational model of decision style can be used as a starting point for the development of computational models to analyze speech data for decision style in other domains. The usefulness of linguistic features supports the applicability of computational modeling to decision style more broadly, since linguistic data may be captured conveniently and non-invasively. Accordingly, future empirical work might focus on modeling and understanding domain influence. This application, of analyzing language data in real time or post hoc, is especially relevant for tasks in which verbal communication is a natural and integral part. This includes team contexts, such as air traffic control, certain medical contexts, and crisis management. If reliable linguistic markers of intuitive and analytical reasoning are in fact uncovered for such contexts, individual or team performance can be measured with respect to the mode of reasoning employed. These results can then be correlated with measures of performance on various tasks in order to determine which type of reasoning best suits particular tasks, as per cognitive continuum theory (Hammond, 2000). Interestingly, such analyses may find that certain tasks can be performed equally well using either type of reasoning; this might then inspire further study of why this is the case. In addition, for those tasks for which expertise is correlated with an increase in intuitive reasoning, linguistic measures can also be used to track individuals as they progress from novice to expert in a certain domain. Finally, since intuitive reasoning has sometimes been associated with biases, it can be detected and extracted from a language database, and then examined for evidence of biases, especially in the case of novice-training contexts. This is especially relevant in contexts in which individuals perform under real or perceived time pressure, which may induce intuitive reasoning (Hamm, 1988).

In addition, the computational model of decision style, after additional research to improve prediction across multiple clinical contexts, may be used in clinical instructional contexts with natural language interfaces. Based on linguistic input, this model can be used to assess whether trainees are using the appropriate style for a particular task (Hammond, 1981), and it can help users determine and attend to their own decision styles, towards improving diagnostic skill (Norman, 2009). This modeling, since it is successful on the basis of only linguistic features, can be useful even when demographic or case difficulty features are unavailable. Such language-based measures of decision style can also be used to assess whether interventions, such as those promoting metacognitive awareness, are effective in promoting flexible and task-appropriate use of Systems 1 and 2. Instructional systems might also track the stability of decision style preferences over time, and also be used to study the effects of metacognitive interventions on diagnostic accuracy and decision style.

Conclusion

This work suggests that decision style is revealed in language use, in line with claims that linguistic data reflect speakers' cognitive processes (Pennebaker & King, 1999). Theoretically, this study adds validity to the dual process and cognitive continuum theories, and articulates a novel way of measuring decision-making style from linguistic data. Methodologically, this study also details strategies for the annotation of fuzzy semantic phenomena and label selection for their modeling, as well as tools to understand annotator strategy. In addition, this work proposed several metrics of decision style at both the case and physician level, based on the developed annotation scale. Analyses based on these metrics found that intuitive reasoning is linked to greater diagnostic accuracy; that experts enjoy greater such accuracy than non-experts when using an intuitive decision style; and that diagnostic difficulty is linked to greater use of analytical decision-making. Meanwhile, analyses regarding the link between metacognitive awareness and decision style were inconclusive, so this relationship deserves future study, both in clinical contexts and in other domains.

Practically, detection of decision style is useful for both clinical educational systems and mission-critical environments. Clinical instructional systems can assess whether participants are using the appropriate style for a particular task (Hammond, 1981), and help students determine and attend to their own decision styles, towards improving diagnostic skill (Norman, 2009). In mission-critical environments, linguistic markers of decision style can be used to determine the optimal modes of reasoning for tasks in high-stakes human factors domains.

References

Allinson, C. W., & Hayes, J. (1996). The cognitive style index: A measure of intuition-analysis for organizational research. *Advances in Applied Sociology, 3*(2), 137-141.

Alm, C. O. (2011, June). Subjective natural language problems: Motivations, applications, characterizations, and implications. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short papers-Volume 2* (pp. 107-112). Association for Computational Linguistics.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*(4), 569-576.

Altman, D. (1991). *Practical statistics for medical research.* London: Chapman and Hall.

Anderson, B., & Shyu, C. R. (2011). Studying visual behaviors from multiple eye tracking features across levels of information representation. In *AMIA Annual Symposium Proceedings 2011* (pp. 72-79). American Medical Informatics Association.

Artz, A. F., & Armour-Thomas, E. (1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. *Cognition and Instruction, 9*(2), 137-175.

Backlund, L., Skånér, Y., Montgomery, H., Bring, J., & Strender, L. E. (2003). Doctors' decision processes in a drug-prescription task: The validity of rating scales and think-aloud reports. *Organizational Behavior and Human Decision Processes, 91*(1), 108-117.

Baker, L., & Brown, A.L. (1984). Metacognitive skills and reading. In P.D. Pearson (Ed.), *Handbook of reading research* (pp. 353-394). New York; Longman.

Bannert, M., & Mengelkamp, C. (2008). Assessment of metacognitive skills by means of instruction to think aloud and reflect when prompted. Does the verbalisation method affect learning? *Metacognition and Learning, 3*(1), 39-58.

Banning, M. (2008). The think aloud approach as an educational tool to develop and assess clinical reasoning in undergraduate students. *Nurse Education Today, 28*(1), 8-14.

Batha, K., & Carroll, M. (2007). Metacognitive training aids decision making. *Australian Journal of Psychology, 59*(2), 64-69.

Baranski, J. V., & Petrusic, W. M. (1998). Probing the locus of confidence judgments: Experiments on the time to determine confidence. *Journal of Experimental Psychology: Human Perception and Performance, 24*, 929-945.

Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., & Rellinger, E. R. (1995). Metacognition and problem solving: A process-oriented approach. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(1), 205-223.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine, 121*, S2-S23.

Beyer, H., & Holtzblatt, K. (1997). *Contextual design: Defining customer-centered systems.* San Francisco, CA: Morgan Kaufmann.

Bornstein, B. H., & Zickafoose, D. J. J, (1999). “I know I know it, I know I saw it”: The stability of the confidence-accuracy relationship across domains. *Journal of Experimental Psychology: Applied, 5*, 76-88.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Brown, A. L. (1978). Knowing when, where, and how to remember: A problem of metacognition. In R. Glaser (Ed.), *Advances in instructional psychology: Vol. 1* (pp. 77-165). Hillsdale, NJ: Erlbaum.

Bullard, J., Alm, C. O., Qi, Y., Shi, P. & Haake, A. (2014, August). Towards multimodal modeling of physicians' diagnostic confidence and self-awareness using medical narratives. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)* (pp. 1718âĂŞ1727). Dublin, Ireland.

Burstein, J., & Chodorow, S. S. M. (2014). Finding your âĂIJinner-annotatorâĂİ: An experiment in annotator independence for rating discourse coherence quality in essays. In *Proceedings of the 8th Linguistic Annotation Workshop at the 25th International Conference on Computational Linguistics* (pp. 48-53). Dublin, Ireland: International Committee on Computational Linguistics.

Cader, R., Campbell, S., & Watson, D. (2005). Cognitive continuum theory in nursing decision-making. *Journal of Advanced Nursing, 49*(4), 397-405.

Coderre, S., Mandin, H., Harasym, P. H., & Fick, G. H. (2003). Diagnostic reasoning strategies and diagnostic success. *Medical Education, 37*(8), 695-703.

Coderre, S., Wright, B., & McLaughlin, K. (2010). To think is good: querying an initial hypothesis reduces diagnostic error in medical students. *Academic Medicine, 85*(7), 1125-1129.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213-220.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15*(10), 687-693.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B, 34*(2), 187-220.

Croskerry, P. (2003a). Cognitive forcing strategies in clinical decision making. *Annals of Emergency Medicine, 41*, 110-120.

Croskerry, P. (2003b). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78*, 775-780.

Croskerry, P. (2006). Critical thinking and decision-making: Avoiding the perils of thin-slicing. *Academic Medicine, 48*(6), 720-722.

Croskerry, P. (2009). A universal model of diagnostic reasoning. *Academic Medicine, 84*(8), 1022-1028.

Croskerry, P., & Norman, G. (2008). Overconfidence in clinical decision making. *The American Journal of Medicine, 121*(5), S24-S29.

Dane, E., Rockmann, K. W., & Pratt, M. G. (2012). When should I trust my gut? Linking domain expertise to intuitive decision-making effectiveness. *Organizational Behavior and Human Decision Processes, 119*(2), 187-194.

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review, 20*(4), 91-409.

Duncker, K.,& Lees, L. S. (1945). On problem-solving. *Psychological monographs, 58*(5), Whole No. 270.

El Saadawi, G. M., Azevedo, R., Castine, M., Payne, V., Medvedeva, O., Tseytlin, E., ... & Crowley, R. S. (2010). Factors affecting feeling-of-knowing in a medical intelligent tutoring system: The role of immediate feedback as a metacognitive scaffold. *Advances in Health Sciences Education, 15*(1), 9-30.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

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

Evans, J. (1989). *Bias in human reasoning: Causes and consequences.* Hillsdale, NJ: Erlbaum.

Evans, J. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454-459.

Evans, J. (2006). The heuristic analytic theory of reasoning: extension and evaluation. *Psychonomic Bulletin and Review, 13*(3), 378-395.

Evans, J. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology, 59*, 255-278.

Evans, J., & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. *Thinking and Reasoning, 11* (4), 382-389.

Ewell-Kumar, A. (1999). The influence of metacognition on managerial hiring decision making: Implications for management development. *Dissertation Abstracts International, A (Humanities and Social Sciences)* (10-A), 3714.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist, 34*, 906-911.

Garofalo, J., & Lester Jr, F. K. (1985). Metacognition, cognitive monitoring, and mathematical performance. *Journal for Research in Mathematics Education,16*(3), 163-176.

Graber, M. L. (2009). Educational strategies to reduce diagnostic error: Can you teach this stuff?. *Advances in Health Sciences Education, 14*, 63-69.

Graber, M. L., Gordon, R., & Franklin, N. (2002). Reducing diagnostic errors in medicine: what's the goal?. *Academic Medicine, 77*(10), 981-992.

Graber, M. L., Kissam, S., Payne, V. L., Meyer, A. N., Sorensen, A., Lenfestey, N., ... & Singh, H. (2012). Cognitive interventions to reduce diagnostic error: A narrative review. *BMJ Quality & Safety, 2*(7), 535-557.

Guo, X., Li, R., Alm, C., Yu, Q., Pelz, J., Shi, P., & Haake, A. (2014, March). Infusing perceptual expertise and domain knowledge into a human-centered image retrieval system: A prototype application. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 275-278). Association for Computing Machinery.

Jakobsson, N., Levin, M., & Kotsadam, A. (2013). Gender and overconfidence: effects of context, gendered stereotypes, and peer group. *Journal of Management Studies, 33*(1), 119-135.

Hacker, D. J. (1998). Definitions and empirical foundations. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 1-24). Mahwah, NJ: Erlbaum.

Hall, K. (2002). Reviewing intuitive decision making and uncertainty: The implications for medical education. *Medical Education, 36*, 216-224.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter, 11*(1), 10-18.

Hamm, R. M. (1988). Clinical intuition and clinical analysis: Expertise and the cognitive continuum. In J. Dowie & A.S. Elstein (Eds.), *Professional judgment: A reader in clinical decision making* (pp. 78-105). Cambridge, England: Cambridge University Press.

Hammond, K. R. (1981). *Principles of organization in intuitive and analytical cognition (Report #231).* Boulder, CO: University of Colorado, Center for Research on Judgment & Policy.

Hammond, K. R. (1996). *Human judgement and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* New York, NY: Oxford University Press.

Hammond, K. R. (2000). *Judgments under stress.* New York, NY: Oxford University Press

Hampton, J. A. (1998). Similarity-based categorization and fuzziness of natural categories. *Cognition, 65*(2), 137-165.

Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences, 1*(2), 78-82.

Hayes, J., Allinson, C. W., & Armstrong, S. J. (2004). Intuition, women managers and gendered stereotypes. *Personnel Review, 33*(4), 403-417.

Hochberg, L., Alm, C. O., Rantanen, E. M., DeLong, C. M., & Haake, A. (2014a). Decision style in a clinical reasoning corpus. In *Proceedings of the BioNLP Workshop* (pp. 83-87). Baltimore, MD: Association for Computational Linguistics.

Hochberg, L., Alm, C. O., Rantanen, E. M., Yu, Q., DeLong, C. M., & Haake, A. (2014b). Towards automatic annotation of clinical decision-making style. In *Proceedings of the 8th Linguistic Annotation Workshop at the 25th International Conference on Computational Linguistics* (pp. 129-138). Dublin, Ireland: International Committee on Computational Linguistics.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics of intuitive judgment: Extensions and applications* (pp. 49-81). New York, NY: Cambridge University Press.

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

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*(1), 23-31.

Katz, J. (1984). Why doctors don't disclose uncertainty. *The Hastings Center Report, 14*(1), 35-44.

Klein, G. A. (1999). *Sources of power: How people make decisions.* Chicago, IL: MIT press.

Kluwe, R. H. (1982). Cognitive knowledge and executive control: Metacognition. In D. R. Griffin (Ed.), *Animal mind – human mind* (pp. 201-224). New York, NY: Springer-Verlag.

Krug, K. (2007). The relationship between confidence and accuracy: Current thoughts of the literature and a new area of research. *Applied Psychology in Criminal Justice, 3*(1), 7-41.

Lauri, S., & Salanterä, S. (1994). Decision-making styles and strategies in nurses' and public health nurses' clinical judgement. In *The Contribution of Nursing Research: Past-Present-Future (Vol. 2)*, 587-597.

Lauri, S., Salanterä, S., Chalmers, K., Ekman, S. L., Kim, H. S., Käppeli, S., & MacLeod, M. (2001). An exploratory study of clinical decision-making in five countries. *Journal of Nursing Scholarship, 33*(1), 83-90.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174.

Li, R., Pelz, J., Shi, P., & Haake, A. R. (2012). Learning Image-Derived Eye Movement Patterns to Characterize Perceptual Expertise. In N. Miyake, D. Peebles, & R.P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 190-195). Austin, TX: Cognitive Science Society.

Lövgren, T., & Doorn, J. V. (2005). Influence of manipulation of short silent pause duration on speech fluency. In *Proceedings of Disfluency in Spontaneous Speech Workshop* (pp. 123-126). International Speech Communication Association.

Lundgrén-Laine, H., & Salanterä, S. (2010). Think-aloud technique and protocol analysis in clinical decision-making research. *Qualitative Health Research, 20*(4) 565-575.

Mamede, S., van Gog, T., van den Berge, K., Rikers, R. M., van Saase, J. L., van Guldener, C., & Schmidt, H. G. (2010). Effect of availability bias and reflective reasoning on diagnostic accuracy among internal medicine residents. *The Journal of the American Medical Association, 304*(11), 1198-1203.

McCoy, W., Alm, C. O., Calvelli, C., Li, R., Pelz, J. B., Shi, P., & Haake, A. (2012, July). Annotation schemes to encode domain knowledge in medical narratives. In *Proceedings of the 6th Linguistic Annotation Workshop* (pp. 95-103). Association for Computational Linguistics.

McCoy, W., Alm, C. O., Calvelli, C., & Pelz, J. B.., Shi, P., & Haake, A. (2012, July). Linking uncertainty in physicians' narratives to diagnostic correctness. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics* (pp. 19-27). Association for Computational Linguistics.

Mokhtari, K., & Reichard, C. A. (2002). Assessing students' metacognitive awareness of reading strategies. *Journal of Educational Psychology, 94*(2), 249-259.

Nielsen, J., Clemmensen, T., & Yssing, C. (2002, October). Getting access to what goes on in people's heads: Reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction* (pp. 101-110). Association for Computational Linguistics.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*, 109-133.

Nelson, T. O. (1996). Gamma is a measure of the accuracy of predicting performance on one item relative to another, not of the absolute performance on an individual item. *Applied Cognitive Psychology, 10*, 257-260.

Neto, A. & Valente, M.O. (1997). Problem solving in physics: Towards a metacognitively developed approach. Paper presented at the *Annual Meeting (70th) of the National Association for Research in Science Teaching.* Oak Brook, IL.

Norman, G. (2009). Dual processing and diagnostic errors. *Advances in Health Sciences Education, 14*(1), 37-49.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1-4.

Paris, S. G., & Oka, E. R. (1986). Children's reading strategies, metacognition, and motivation. *Developmental Review, 6*(1), 25-56.

Paris, S. G., & Winograd, P. (1990). How metacognition can promote academic learning and instruction. In B.F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp.15-51). Hillsdale, NJ: Lawrence Erlbaum Associates.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296-1312.

Pintrich, P. R, & De Groot, E. (1990). Motivational and self-regulated learning components of classroom academic performance. *Journal of Educational Psychology, 82*(1), 33-40.

Pintrich, P. R., Wolters, C. A., & Baxter, G. P. (2000). Assessing metacognition and self-regulated learning. In J. C. Impara, G. Schraw, & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 43-97). Lincoln, NE: University of Nebraska-Lincoln.

Prime, N. J., & Le Masurier, S. B. (2000). Defining how we think: an investigation of decision making processes in diagnostic radiographers using the 'think aloud' technique. *Radiography, 6*(3), 169-178.

Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning.* Sebastopol, CA: O'Reilly Media, Inc.

Quirk, M. E. (2006). *Intuition and metacognition in medical education: Keys to developing expertise.* New York, NY: Springer Publishing Company.

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica, 127*, 258-276.

Sadler-Smith, E. (2011). The intuitive style: Relationships with local/global and verbal/visual styles, gender, and superstitious reasoning. *Learning and Individual Differences, 21*(3), 263-270.

Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring *Metacognition Learning, 4*, 33-56.

Shanteau, J. (1988). Psychological characteristics and strategies of expert decision makers. *Acta Psychologica, 68*, 203-215.

Sherbino, J., Dore, K. L., Siu, E., & Norman, G. R. (2003a). (2011). The effectiveness of cognitive forcing strategies to decrease diagnostic error: An exploratory study. *Teaching and Learning in Medicine, 23*(1), 78-84.

Sjöberg, L. (2003). Intuitive vs. analytical decision making: Which is preferred? *Scandinavian Journal of Management, 19*(1), 17-29.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences, 23,* 645-665.

Sternberg, R. J. (1998). Metacognition, abilities, and developing expertise: What makes an expert student?. *Instructional science, 26*, 127-140.

Swanson, H. L. (1990). Influence of metacognitive knowledge and aptitude on problem solving. *Journal of Educational Psychology, 82*(2), 306-314.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54.

Tobias, S., & Everson, H. (1995, April). Development and validation of an objective measure of metacognition. In W. E. Montague (Chair), *Issues in metacognitive research and assessment.* San Franciso, CA: American Educational Research Association.

Thompson, V. A. (2009). Dual process theories: A metacognitive perspective. In J. Evans and K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 171-195). Oxford, UK: Oxford University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology, 88*(3), 490-499.

Weinstein, C. E., Schulte, A, & Palmer, D. (1987). *LASSI: Learning and Study Strategies Inventory.* Clearwater, FL: H&H Publishing.

Womack, K., McCoy, W., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, P., & Haake, A. (2012, July). Disfluencies as extra-propositional indicators of cognitive processing. *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics* (pp. 1-9). Association for Computational Linguistics.

Womack, K., Alm, C.O., Calvelli, C., Pelz, J.B., Shi, P., & Haake, A. (2013, August). Markers of confidence and correctness in spoken medical narratives. In *Proceedings of Interspeech 2013* (pp. 2549-2553). International Speech Communication Association.

Womack, K., Alm, C. O., Calvelli, C., Pelz, J. B., Shi, P., & Haake, A. (2013, August). Using linguistic analysis to characterize conceptual units of thought in spoken medical narratives. In *Proceedings of Interspeech 2013* (pp. 3722-3726). International Speech Communication Association.

Zimmerman, B. J., & Martinez-Pons, M. (1986). Development of a structured interview for assessing student use of self-regulated learning strategies. *American Educational Research Journal, 23*, 614-628.

Appendix A

Concept Map of Study Variables and Hypotheses

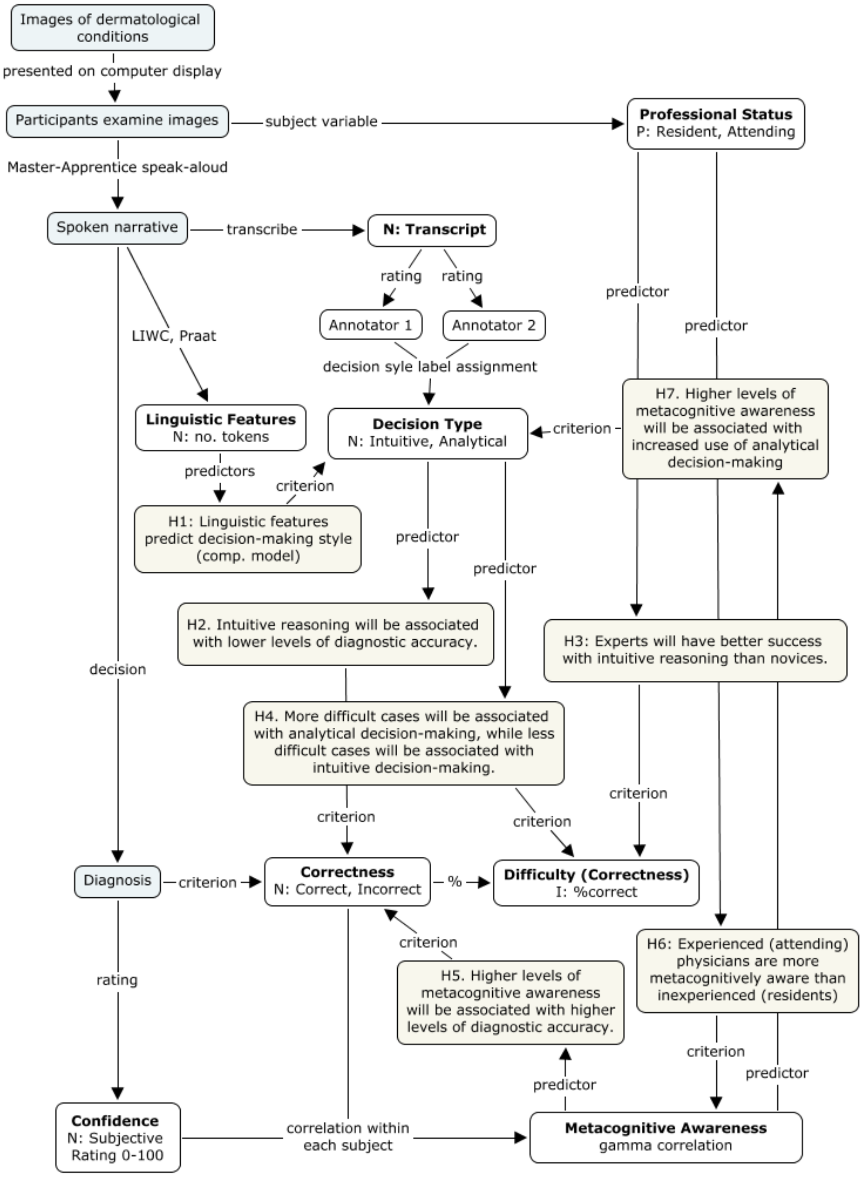

Images of dermatological conditions
presented on computer display
Participants examine images
subject variable
Professional Status
P: Resident, Attending
Master-Apprentice speak-aloud
Spoken narrative
transcribe
N: Transcript
rating
rating
Annotator 1
Annotator 2
predictor
predictor
LIWC, Praat
decision syle label assignment
Linguistic Features
N: no. tokens
Decision Type
N: Intuitive, Analytical
criterion
H7. Higher levels of metacognitive awareness will be associated with increased use of analytical decision-making
predictors
criterion
H1: Linguistic features predict decision-making style (comp. model)
predictor
predictor
H2. Intuitive reasoning will be associated with lower levels of diagnostic accuracy.
H3: Experts will have better success with intuitive reasoning than novices.
decision
H4. More difficult cases will be associated with analytical decision-making, while less difficult cases will be associated with intuitive decision-making.
criterion
criterion
criterion
Diagnosis
criterion
Correctness
N: Correct, Incorrect
%
Difficulty (Correctness)
I: %correct
criterion
H6: Experienced (attending) physicians are more metacognitively aware than inexperienced (residents)
rating
H5. Higher levels of metacognitive awareness will be associated with higher levels of diagnostic accuracy.
criterion
predictor
predictor
Confidence
N: Subjective Rating 0-100
correlation within each subject
Metacognitive Awareness
gamma correlation

Figure A1. This concept map illustrates the major study variables and hypotheses.

Appendix B

Instructions to Annotators

Task Description

We conducted an experiment in which dermatologists were asked to look at digital images of patients with skin disorders and describe the findings to a trainee while working towards the diagnoses.

The attached documents are the transcribed versions of the verbal descriptions captured during the above experiment. Since they are verbatim transcripts of the audio they might contain repeated words, incomplete sentences, disfluencies like *uh* and *um*, etc.

New to instructions since pilot:

- Narratives now include ellipses, which indicate pauses
- Indications of physician confidence have been removed from the narratives, either by removal of final clauses containing the confidence judgments or by substituting dashes for certain words (for example "—– percent certainty").
- Five questionnaires have been added – please answer them as you encounter them based on the last set of narratives rated; please let me know if you have any questions!
- Please also note that diagnoses are sometimes abbreviated, such as *iga* or *scle*.

We would like you to rate each narrative on its predominant decision-making style.

Rating Scale

Please rate each narrative as:

I – *Intuitive*– reflecting primarily intuitive processes and decision-making

BI – *Both, but intuitive appears more dominant* – reflecting intermediate values of the features; reflecting a mix of characteristics, from both intuitive or analytical modes; or oscillating between the two modes; but with more tendency towards the intuitive style
BA – *Both, but analytical appears more dominant* – reflecting intermediate values of the features; reflecting a mix of characteristics, from both intuitive or analytical modes; or oscillating between the two modes; but with more tendency towards the analytical style

A – *Analytical* – reflecting primarily analytical processes and decision
Please put the letter code in the "Rating" column to the right of each narrative.

Features of Intuitive or Analytical Decision-making

Intuitive decision-making is:

- holistic – considers the case as a whole
- based on simultaneous use of cues
- automatic – involuntary; requires little or no conscious effort and attention
- associative – draws connections between related ideas in a non-linear fashion

Analytical decision-making is:

- step-by-step
- based on sequential use of cues
- governed by rules of logic and domain principles
- likely to include justification

Please use these guidelines while rating the narratives, as well as your general intuitions based on your previous knowledge of human factors and cognitive psychology.

Examples

You'll notice that the intuitive narratives tend to be shorter than the analytical narratives. *However, please avoid using length in coding the narratives!*

Example of an Intuitive narrative:

... um ... numerous tan to ... gray-brown ... um ... verrucous stuck-on plaques ... differential diagnosis ... seborrheic keratosis ... diagnosis seborrheic keratosis

Example of a Both-Intuitive narrative:

... mm kay s- ... so have ... an axilla that's got ... lots of uh ... redness and ... uh ... looks like swelling ... and obviously uh ... uh flaccid bullae ... with uh ... yellow ... uh ... filling of the bullae ... multiple small ... vesicles ... mostly this looks like a contact dermatitis ... um ... uh could also be bullous pemphigoid pemphigus ... uh hailey-hailey ... um could be those things but looks mostly like ... uh contact dermatitis just with how uh ... um ... large red and angry and ... it has uh some areas of sparing

Example of a Both-Analytic narrative:

... s- is a lower extremity ... extensor aspect ... and there are ... uh ... at least ... ten plus ... um ... small to large fluid-filled ... tense bullae ... so this is a blistering disease ... uh this ... could be pemphigus vulgaris ... could be ... bullous pemphigoid ... um ... most likely is a bullous drug eruption ... or bullous bite reaction ... mm there look to be areas that have healed as well with some early scarring so ... i would say it is one of the bullous diseases ... um ... would definitely have biopsy ... diagnosis ... and ... mm not that old ... um ... i will lean towards ... pemphigoid

Example of an Analytical narrative:

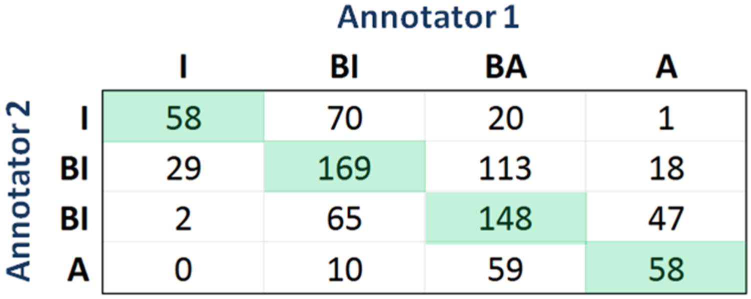
... okay uh i am seeing uh ... a ... polycyclic um ... eruptions uh on ... what appears to be a leg or a thigh ... that is with some pretty intense erythema ... uh also with um ... uh some centrally located vesicles um ... could be a bullae that or at least there's a few little erosions ... um ... and ... uh there's a number of different things that could cause this i would be thinking about like a ... possibly a linear iga bullous dermatosis uh even tinea can do something like this uh ... if it was more uh although i'd expect it to be a little bit more scaly and not sort of uh ... and more centrally located uh ... or a centrally located scale ... uh if it was bullous tinea ... um ... possibly like a ... uh ... eac but uh i uh think if i had to narrow it down i would probably favor like a linear iga bullous dermatosis uh ... or other ... um ... i guess you could think about like uh ... uh bullous lupus potentially um ... for this i'd have to ... i mean i'd really want a biopsy with uh ... for a tinea and for dif and ... i don't know as far as percent certainty i'd say like ... —— percent without some confirmatory evidence ... next ...

Thanks

Thank you for your time. Let me know if you have any questions as you go along!

Appendix C

Decision Style Annotation Confusion Matrix



		Annotator 1			
		I	BI	BA	A
Annotator 2	I	58	70	20	1
	BI	29	169	113	18
	BI	2	65	148	47
	A	0	10	59	58

Figure C1. Confusion matrix for two annotators on the 4-point decision rating scale; full agreement is shown in green. *I = Intuitive, BI = Both-Intuitive, BA = Both-Analytical, A = Analytical.*

Appendix D

Annotator Strategy Questionnaire

Table D1

Annotator Strategy Questionnaire

Questionnaire	**Answer**	
Please respond based on your judgments for the last set of narratives (i.e., since the last questionnaire)		
FREE FORM: How are you rating each narrative?		
What indicates an intuitive (I) style of reasoning?		
What indicates an analytical (A) style of reasoning?		
How do you distinguish between A and BA?		
How do you distinguish between BA and BI?		
How do you distinguish between BI and I?		
How often do you use each factor in rating the narratives?	**I use this factor to rate....**	**Comments**
	1 (no narratives) 2 (few narratives) 3 (about half of the narratives) 4 (most narratives) 5 (all narratives)	
Automatic v. Controlled Processing		
Holistic v. Sequential Processing		
Degree of Associative Processing		
Use of Justification		
Use of Logical Rules and Inference		
Word Choice		Did you use any specific words or phrases?
Silent pauses (...)		
Filled pauses (e.g. uh, um)		
False starts (participant starts a word or phrase, then partially repeats or re-starts)		
Number of diagnoses included in differential diagnosis		
Timing of differential diagnosis		
Relationship between final diagnosis and first or second diagnosis mentioned		

Five identical such questionnaires were presented to annotators, within the Excel document used for decision style annotation.

Table D2

Annotator Strategy Questionnaire - Continued

Relationship between final diagnosis and differential diagnosis		
Whether a participant switched from one more of reasoning to another		
Timing of switch between modes of reasoning		
Whether the differential diagnosis seemed authentic/natural or provided just to fulfill task requirements		
Perceived attitude		
Perceived confidence		
Other macro-level (narrative-level) judgments: please detail		
Other		*please detail here*
Other		*please detail here*
Open-Ended Questions		
Please detail any changes or adjustments to your coding system since the last questionnaire		
Please describe any other general comments and/or concerns		

Appendix E

Complete List of Lexical Features In Computational Model of Decision Style

Table E1

Complete List of Lexical Features Used For Decision Style Modeling

Feature	Examples
function	*at, most, very*
pronoun	*i, she, him*
personal pronoun	*i, them, her*
first person singular pronoun	*i, me, mine*
first person plural pronoun	*we, us, our*
second person pronoun	*you, your*
third person singular	*she, her*
third person plural	*they, their*
impersonal pronoun	*it, those*
article	*a, an, the*
common verbs	*walk, went*
auxiliary verbs	*am, will, have*
future	*will, gonna*
preposition	*to, with, above*
conjunction	*and, but*
negation	*no, not, wasn't*
quantifier	*all, less*
swear	*hell, darn*
affect	*happy, fear*
cognitive process	*know, whether*
insight	*think, know*
cause	*because, therefore*
discrepancy	*could, would*
tentative	*maybe, perhaps*
certainty	*always, never*
inclusion	*both, with*
exclusion	*but, without*
perceptual processes	*feel, hear, press*

Lexical features correspond to categories in the Linguistic Inquiry and Word Count Software (Tausczik & Pennebaker, 2010).

Table E2

Complete List of Lexical Features Used For Decision Style Modeling - Continued

Feature	Examples
see	*look, saw*
biological processes	*eat, blood, pain*
body	*ears, skin*
health	*acne, insulin*
relativity	*following, again*
motion	*go, appear*
space	*farther, underneath*
time	*before, until*
assent	*ok, okay, alright*

Appendix F

Gender Effects on Decision Style and Metacognitive Awareness

This study included two main demographic variables, expertise and gender. In this appendix, the relationship between gender and decision style, diagnostic accuracy, and metacognitive awareness is explored. There were 16 female and 13 male physician participants.

Figure F1 shows physician decision scores by gender (see Equation 3 and *Physician Profiles of Decision Style*, above). A Mann-Whitney test found no significant difference between the male (*Mdn* = 2.9) and female (*Mdn* = 2.4) physicians, U=64, p = .082.

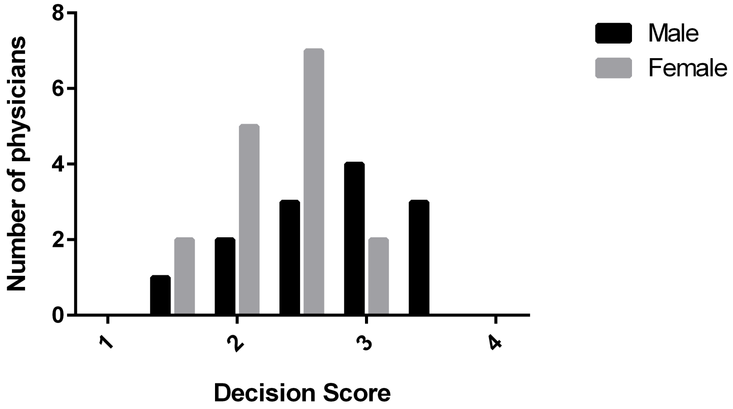


Figure F1. Distribution of physician decision scores by gender.

Figure F2 compares diagnostic accuracy, as measured for each physician by percent correct across all narratives, among male (*Mdn* = .400) and female (*Mdn* = .433) physicians. A Mann-Whitney test revealed no significant difference between the two distributions, U=85.5, p = .423.

Figure F3 compares success in intuitive reasoning (as measured by the intuitive-correct proportion; see Table 9, above) among male (*Mdn* = .467) and female (*Mdn* = .500) physicians. A Mann-Whitney test revealed no significant difference between the two distributions, U=97.5, p = .786.

Figure F4 shows the distribution of metacognitive awareness, as measured by the gamma confidence-accuracy correlation (see *Physician Profiles of Metacognitive Awareness,*

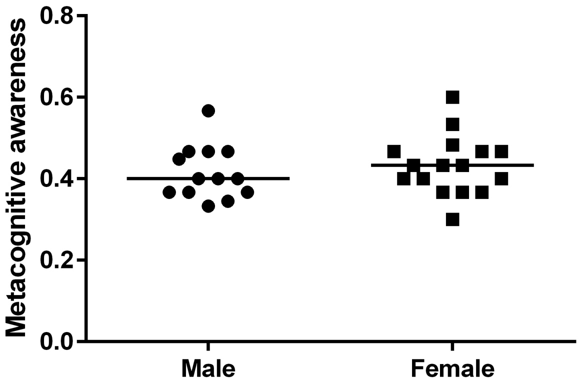


Figure F2. Diagnostic accuracy by gender. The horizontal line indicates the median of each distribution.

above) among male (*Mdn* = .600) and female (*Mdn* = .507) physicians. A Mann-Whitney test revealed no significant difference between the two distributions, U=97, p = .770.

Based on these analyses, no gender effects were found on decision style, diagnostic accuracy, or metacognitive awareness. These results are in line with the computational model developed for automatic annotation of decision style, in that gender was not among the best features for modeling. It may be the case that there are gender effects, but the sample size here was too small to detect them; or, alternatively, that there are not

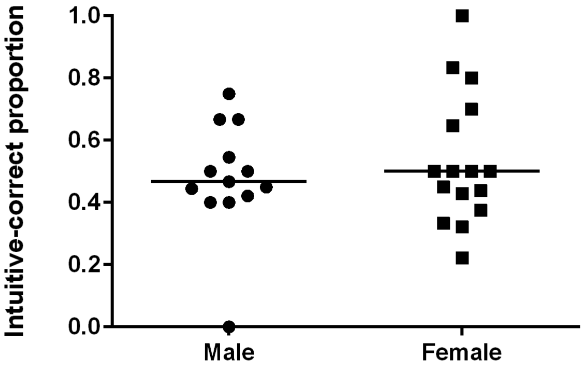


Figure F3. Success in using intuitive reasoning, by gender. The horizontal line indicates the median of each distribution.

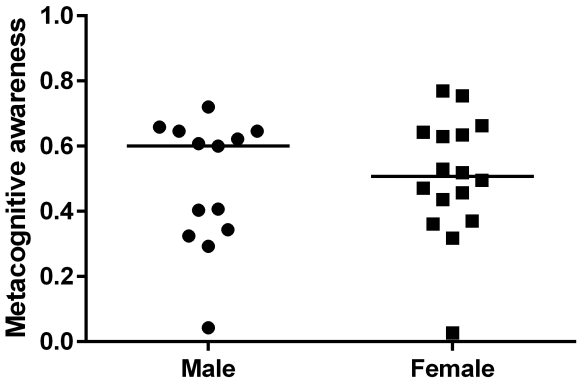


Figure F4. Metacognitive awareness by gender. The horizontal line indicates the median of each distribution.

gender effects on these variables. Future work might investigate this link more carefully, particularly with respect to the link between decision style and gender, which so far in the literature exhibits mixed results (e.g., Hayes, Allinson, & Armstrong, 2004; Sadler-Smith, 2011). In addition, metacognitive awareness might be studied with an eye towards previously reported gender effects on confidence, by which women tend to be less confident than men (see Jakobsson, Levin, & Kotsadam, 2013, for a review).