A conceptual framework to guide the design and evaluation of human interaction with information automation to support judgment

A dissertation

presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

In partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Systems Engineering

by

Leigh A. Baumgart

August, 2012

Approval Sheet

This dissertation is submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Systems Engineering

Leigh A. Baumgart
Author

The dissertation has been read and approved by the examining committee:

Dr. Gregory J. Gerling
Dissertation Advisor

Dr. James H. Harrison
Dissertation Advisor

Dr. Stephen D. Patek
Committee Chair

Stephanie Guerlain
Dr. Stephanie Guerlain

Jason Lyman
Dr. Jason A. Lyman

Accepted for the School of Engineering and Applied Science:

J. H. Aylor
Dean, School of Engineering and Applied Science
August 2012

Abstract

In critical domains from air traffic control to health care, humans make judgments by using informational cues to assess the true state of the environment. How humans judge environmental conditions is now ubiquitously supported by automation. While prior modeling and analysis has defined levels of automated decision support and methods to evaluate independent human and automated judges, we have yet to develop ways to both support and evaluate human judgment.

To address this gap, this research develops a conceptual design and evaluation framework, titled the Expanded Lens Model with Automation (ELMA). To provide design support, ELMA accounts for discrepancies between how cues in the environment are transformed into operator displays via automated processes. The transformation is based upon the desired, hierarchical level of cognitive judgment support, from cue perception, to cue comprehension, to an automated assessment, to explanation of the automated assessment. In addition to design support, ELMA includes quantitative evaluation measures, using multiple linear regression and correlation analysis to characterize the achievement, consistency, and task knowledge of the human judge; potential and accuracy of automation; and predictability of the environment.

ELMA's utility is demonstrated through the investigation of two tasks: 1) judging the probability of an air traffic conflict using heading and speed cues and 2) judging the quality of population-based hypertension care using cues on patient outcomes (e.g., blood pressure) and processes (e.g., medications prescribed). Across both tasks, ELMA revealed that when participants were supported at the cognitive level of cue comprehension they achieved significantly higher judgment achievement compared to those supported at the lower level of cue perception. Decomposition of achievement indicated that these differences were predominantly due to the consistency of individuals in executing judgments rather than task knowledge. Additionally, reliability across participants was significantly higher for participants with cue perception and cue comprehension support compared to participants with automated assessment support in the quality of hypertension care task.

ELMA is a useful tool for systems engineers as it provides both a systematic framework to inform automation design choices and a quantitative method to evaluate human-automation judgment systems.

Acknowledgments

This work was financially supported by Grant Number T15LM009462 from the National Library of Medicine and Research Grant Agreement UVA-03-01. The content is my sole responsibility and does not necessarily represent the views of the National Library of Medicine or the National Institutes of Health.

Additionally, I am fortunate to have had academic and emotional support from several people throughout my PhD program and for them, I am extremely grateful.

First, I would like to thank Dr. Carlos Comperatore and Dr. Jennifer Ockerman for their initial encouragement to pursue a PhD. Their early confidence in me has played an instrumental role in my career and they continue to be two researchers who I greatly respect and admire.

Second, I would like to acknowledge Dr. Ellen Bass. Her expertise and participation in many discussions played a valuable role in the conceptualization and refinement of the framework presented in Chapter 3. I also appreciate her thorough review of that chapter and her providing insights that improved this work. Additionally, I thank both Dr. Bass and Katie Klein Shepley for building the apparatus, collecting the data, and reviewing the manuscript related to the work I present in Chapter 4.

The study presented in Chapter 5 would not have been possible without the tremendous support of both Dr. Jason Lyman and Dr. John Voss. I thank them for patiently listening to my research ideas and for believing that the study was worthwhile. I appreciate their contributions to the conceptualization of the experimental design, designing the apparatus, and interpreting the data analysis. Dr. Voss was also instrumental in facilitating data collection for this study. His support and advocacy made it easy to solicit the cooperation of attending and resident physicians who enthusiastically participated in multiple focus groups and study sessions. For this, I am very grateful.

I am also extremely thankful for the unwavering support and guidance provided by Dr. Gregory Gerling. I appreciate having had the opportunity to work with him on an initial project that laid the foundation for further investigation into the modeling and analysis of human judgment. I also appreciate the time he spent editing multiple iterations of this dissertation and for his very thoughtful feedback. Without a doubt, he has improved my ability to articulate research ideas. I also thank him for always having an open door and for always believing in me.

I sincerely thank my other committee members for their assistance throughout the development of this dissertation. I am grateful to Dr. James Harrison, Dr. Stephen Patek, and Dr. Stephanie Guerlain for serving on my committee, reviewing my dissertation, providing useful feedback, and participating in enjoyable discussions related to my research.

There are also many colleagues and friends who have influenced this research in many ways and have provided immeasurable support and encouragement, especially in the final months of my program. There are more than I can list, but I would particularly like to thank Jenny Prey, Akilah Hugine, Don Rude, Justin DeVoge, and Matthew Bolton.

I would also like to thank my family, especially my parents, brother, and sister. Although, they were not always certain how I was spending my time in school all these years, I knew that they always were (and always will be) very proud of me. Knowing this means very much to me.

Finally, I would like to thank my fiancée, Nora Decher. I am especially appreciative of her willingness to not only listen, but to really want to understand every good and bad idea that has ever entered my brain over the past few years. This includes the many times I described different quantitative models of human performance to her in the middle of long runs on the Rivanna Trail. I am forever indebted to her for her unrelenting support and endless belief in me. I thank her for intimately and affectionately sharing this journey with me.

Table of Contents

List of Figures

List of Tables

1. Overview and objectives

In critical domains from air traffic control to health care, humans make critical judgments. To inform a judgment, one uses available information (or cues) with the goal of formulating an assessment about the true state of the environment before making a decision. For example, an air traffic controller uses aircraft position, speed, and heading cues to judge the likelihood of a collision before deciding to reroute an aircraft. A clinician uses cues indicating the proportion of patients up to date on recommended cancer screenings (e.g., mammography) to judge the quality of preventative health care provided to a group of patients before deciding to allocate resources to a quality improvement initiative.

How humans judge environmental conditions is now ubiquitously supported by automation. We need to understand the interaction of human and automation to inform automation design choices. One approach to studying this interaction is using qualitative frameworks to distinguish types and levels of automation. However, the focus of this approach has been to support decision making and supervisory tasks, instead of supporting cue perception, cue comprehension, and assessment [1–6]. Another approach, which unlike the former that includes subjective evaluation measures, is the use of judgment analysis to decompose achievement into other quantitative measures that reflect consistency, task knowledge, and environmental predictability. However, this approach has focused on evaluating independent human and automated judges or in characterizing how a human adapts to an automated judgment [7–13].

To summarize, we have yet to define levels of automated judgment support and quantitative evaluation measures for human-automation judgment systems. The central hypothesis of the work herein is that we can develop a conceptual design and evaluation framework for automation that supports cognitive functions involved in judgment. This central hypothesis is supported by three aims.

The first aim is to develop the framework, titled the Expanded Lens Model with Automation (ELMA). To provide design support, ELMA accounts for discrepancies between how cues in the environment are transformed into displays to operators via automated processes. The transformation is based upon the desired, hierarchical level of cognitive judgment support, from cue perception, to cue comprehension, to an automated assessment, to explanation of the automated assessment. In addition to design support, ELMA also includes quantitative measures to evaluate the human-automation system with an idiographic-statistical[1] approach. Multiple linear regression and correlation analysis are employed to characterize achievement, consistency, and knowledge of the human judge, potential and accuracy of automation support, and predictability of the environment. Nomothetic analysis can then be applied to investigate the effect of automation design choices on general judgment performance.

The second aim is to demonstrate ELMA's utility and address domain-specific objectives, by evaluating an existing system that supports air traffic controllers in *judging the probability of an air traffic conflict* using heading and speed cues. This effort is significant because the Federal Aviation Administration (FAA) predicts that air traffic will more than double in the next twenty years [14]. This growth emphasizes the need to systematically and quantitatively evaluate the impact of automation design choices on conflict judgments prior to incurring the high costs of full-scale development, deployment, training, and maintenance. Specifically, the objectives of this aim were to investigate the effect of providing additional levels of support, in tandem with an automated judgment, on operator achievement and consistency. ELMA revealed that participants provided with cue comprehension support had significantly higher judgment achievement compared to those provided with only automated assessment support. This was predominantly due to improved consistency, in that participants may have been able to better understand the environmental context and were thus able to more consistently apply their judgment strategies.

The third aim is to demonstrate ELMA's utility and to understand and inform the design of automation to support physicians in *judging the quality of population-based hypertension care* using

[1] An idiographic-statistical approach investigates individuals before general (nomothetic) trends.

cues on patient outcomes (e.g., blood pressure) and processes (e.g., medications prescribed). This effort is significant because resident physicians are now required to demonstrate the ability to evaluate the quality of their care as a core learning requirement instituted by the Accreditation Council for Graduate Medical Education (ACGME) [15]. Yet, physicians have shown a limited ability to accurately make such judgments [16] and few automated tools are available to support them [17]. Specifically, the objectives of this aim were: 1) to identify cues needed to judge hypertension care quality, 2) to understand how these cues differentially influence judgment, and 3) to evaluate the impact of level of automated support on achievement, consistency within individual physicians, and reliability across physicians. ELMA revealed that the percentage of patients at goal blood pressure had the greatest impact on quality judgments compared to the percentage of patients on recommended medications and with recommended laboratory tests checked. Further, resident physicians supported with cue perception and cue comprehension had significantly better achievement, consistency, and reliability compared to those supported with an automated assessment.

2. Conceptual background

This chapter begins by defining human judgment. This is followed by a description of frameworks used to design and evaluate automation systems, which has motivated the need for the ELMA framework to guide design to specifically support human judgment. A discussion of conceptual, quantitative models to investigate human judgment, predominantly those arising from the lens model-based approach are also included. These models have inspired the conceptualization of ELMA. Prior empirical findings related to automation-supported judgment performance are also included. This chapter concludes with a summary of the conceptual gaps in the literature and outlines the research objectives of this dissertation.

2.1. Human judgment

To make a judgment, a judge considers one or more cues. Cues represent any information elements available to the judge and they can be gathered from any sensory input (e.g., visual, aural, haptic). The judge perceives, comprehends, and integrates the cues to formulate a judgment regarding the true state of the environment, where the true state is known as the environmental criterion. This process is also referred to as the “front-end judgment process.” This is differentiated from the “back-end decision process” that, based on the judgment, involves generating possible courses of action, weighing one’s options, and mentally simulating a possible response [18].

For example, a naval officer uses multiple cues to make a judgment about the identity of an aircraft approaching the officer’s ship [19]. The cues might include the speed, altitude, and range of the approaching aircraft, informed by a visual user interface. The environmental criterion would be the true identity of the aircraft (e.g., hostile or friendly). Each of the cues used to make the judgment also have a direct relationship to the environmental criterion. For example, aircraft that fly at speeds less than 600 knots are typically friendly (e.g., commercial). Once the officer makes a judgment about the identity of the aircraft, he or she may then decide whether or not to make contact with the aircraft. This research focuses on the front-end judgment process, rather than the back-end decision process that may follow or be based on a judgment.

2.2. Frameworks for automation design and evaluation

Automated systems are increasingly being designed to assist humans in making judgments in domains such as defense [19], weather [20], aviation [21], and health care [17], [22], [23]. In the domain of health care, for example, such systems are expected to proliferate given the Institute of Medicine's promotion of health information technology [24] and the push for widespread adoption of information technology through the Health Information Technology for Economic and Clinical Health Act (HITECH) [25].

Automation may gather and present cues from the environment, provide an evaluation of cues, integrate cues to formulate an automated judgment, or explain its own cue integration strategy for the human to draw on. When designing automation to support human judgment, both the functionality (i.e., content of information presented) and the representation of information must be considered. This section will review existing frameworks related to automation functionality (the focus of this dissertation). For reviews on designing the representation of information, see [26–28].

2.2.1. Function allocation frameworks

Traditionally, automation designers made decisions about what cognitive functions to automate in a qualitative manner by simply automating whatever was possible from a technical and/or cost perspective. As technical developments improved, the main automation design question became: *what system (i.e., cognitive) functions should be automated and to what extent* [1]*?*

Fitts et al. [29] were the first to suggest a framework to guide decisions on *what* to automate based on function allocation methods. Designers were encouraged to automate all tasks where machines perform better than humans. For example, it was thought that men were better at perceiving patterns of light or sound and machines were better at storing information briefly and then erasing it completely [30]. Similar MABA-MABA lists, or ‘Men are Better At-Machines are Better At,’ have been proposed in different domains (e.g., [31][32]). Others have attempted to use these function allocation methods in order to guide automation design [29], [33], [34], but they have been found to be difficult to use [32], [35]. Further, function allocation does not necessarily mean allocation of a whole task to either human or machine, exclusive of the other [30]. This concept has inspired the development of frameworks that incorporate taxonomies to describe levels or degrees of automation for a particular function [2–5], [36]. One commonly cited taxonomy was developed by Sheridan and Verplank [36]. It consists of ten levels of automation as depicted in Table 1.

Table 1. Sheridan and Verplank's levels of automation [36].

Level of Automation
1. Human does the whole job up to the point of turning it over to the computer to implement.
2. Computer helps by determining the options.
3. Computer helps to determine options and suggests one, which the human need not follow.
4. Computer selects an action and the human may or may not do it.
5. Computer selects an action and implements it if the human approves.
6. Computer selects an action and informs the human in plenty of time to stop it.
7. Computer does the whole job and necessarily tells the human what it did.
8. Computer does the whole job and tells human what it did only if the human explicitly asks.
9. Computer does the whole job and decides what the human should be told.
10. Computer does the whole job if it decides it should be done, and if so, tells the human, if it decides that the human should be told.

While this taxonomy can guide design in some contexts, it is limited to situations involving the sharing of functions related to determining options, selecting options, and implementing options. Others have expanded this work to address a wider array of cognitive and psychomotor tasks. In another framework, ten levels of automation were formulated by assigning four different cognitive functions (monitoring, generating, selecting, and implementing) to the human, the automation or both [2–5]. This taxonomy provides a systematic way to examine the effect of automation applied to support more cognitive functions, as implemented incrementally. However, it was developed for the design of automation to support supervisory control tasks and would be difficult to apply to the design of automation to support judgment tasks where the goal is to make an accurate assessment of the environment.

2.2.2. Parasuraman, Sheridan, and Wickens (PSW) framework for "types" and "levels" of automation

Parasuraman et al. [1] suggest that to guide the design of automation functionality to support an array of cognitive tasks, a simple four-stage view of human information processing is first derived from a more detailed model [37] (Figure 1).

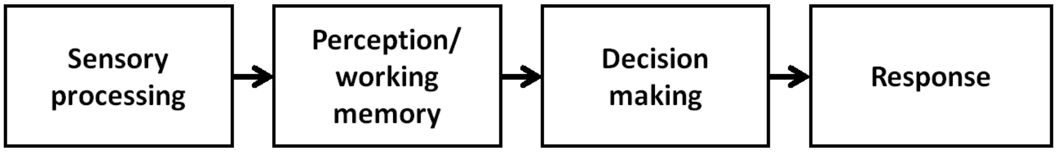


Figure 1. Simple four-stage model of human information processing [1].

The first stage of information processing refers to the acquisition of information (or cues) from the environment. The second stage represents conscious perception and manipulation of retrieved information in working memory to form an assessment, or to make a judgment regarding the state of the environment. The third stage is where alternative decisions are generated and one is selected and the fourth stage involves implementation of a response or action consistent with the decision [1]. Although, the authors acknowledge that this is a gross simplification of human information processing, similar conceptual models have been found useful for design decisions for other systems [38].

This four stage model then has an equivalent in system functions that can be automated to support the human's information processing stages. Four "types" of automation are defined as indicated in Figure 2. Information acquisition automation includes functionality to sense and register input data from the environment. This could include automation designed to gather and organize incoming cues for the human judge. Information analysis automation includes functions related to working memory, integration, or inferential processing to make an assessment or judgment about the environment. This type of automation could include algorithms applied to categorize or evaluate cues or algorithms used to integrate cues to automatically judge the current state of the environment.

Together, the first two types of automation are often referred to as *Information Automation (IA)* and they most directly support the cognitive functions involved in human judgment. The first two types of automation have also been suggested as possible interventions to benefit situation awareness (SA) [8], [10]. The SA construct, which was postulated by Endsley [39], includes the perception of cues in the environment, comprehension of the current situation, and the projection of the current situation into the near future.

Decision and action selection automation (the third type) includes functions such as augmenting or replacing the way a human generates and selects between decision alternatives. The final type of automation, action implementation, is designed to execute the decision or choice of action.

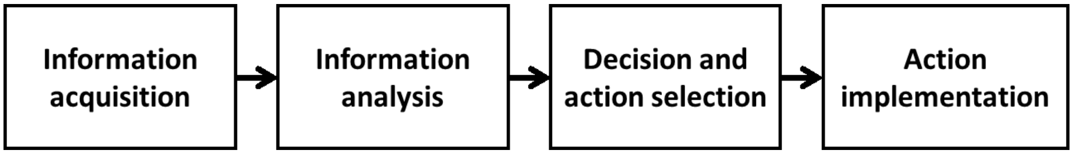


Figure 2. Four-stage model of automation types.

Each automation "type" can then be designed to provide a different "level" of support, creating a two-dimensional framework of types and levels to guide the design of automation. However, Parasuraman et al. [1] only suggest "levels" for one of the three "types" of automation: the decision and action selection type (Table 2).

Table 2. Levels of automation for the decision and action selection type.

	Level of Automation
High 10.	The automation decides everything, acts autonomously, ignoring the human.
9.	The automation informs the human only if it decides to.
8.	The automation informs the human only if asked.
7.	The automation executes automatically and then necessarily informs the human.
6.	The automation allows the human a restricted time to veto before automatic execution.
5.	The automation executes the suggestion if the human approves.
4.	The automation suggests one alternative.
3.	The automation narrows the selection down to a few alternatives.
2.	The automation offers a complete set of alternatives.
Low 1.	The automation offers no assistance; human must make all decisions and actions.

Given the types and levels of automation suggested by Parasuraman et al. [1], [6] a series of iterative steps is prescribed for their integration into a conceptual framework for making automation design choices and then evaluating the human-automation system. The first step is to pick the type(s) of automation. The second step is to choose the desired level of automation per type. (However, no guidance is provided for this step and it is even suggested that there is “no simple answer.”) The authors propose that any type/level combination should then be evaluated by examining the associated human performance consequences. These include investigating mental workload, situation awareness, complacency, and skill degradation [1].

The objective of this framework is to support automation design decisions on the basis of empirical evaluation of various combinations of the types and levels. This framework provides the most comprehensive guide for automation design to support human information processing to date (particularly in the design of automation to support human decision making). However, it can essentially only be used as a starting point in the design of automation to support human judgment (i.e., IA) because no levels of IA have been defined.

Further, while quantitative models of human performance have been suggested to supplement this qualitative framework, no model has been suggested to support the design of IA (specifically at the analysis stage) [6]. Quantitative models of human judgment may aid in the analysis of benefits and costs associated with different levels of IA. The following section will review conceptual, quantitative models of human judgment (corresponding with the first two stages of human information processing as indicated in Figure 1) that may fit this need.

2.3. Conceptual models to investigate human judgment

While many different conceptual models have been developed to investigate human judgment, few specifically address the relationships between the cues, the human judge, and the environmental criterion. For example, of the fourteen judgment and decision making approaches analyzed by Cooksey [40], only signal detection theory (SDT) and judgment analysis (JA) consider these relationships. The focus of SDT is on discrimination (i.e., a judge's ability to detect a signal against noise), only one aspect of the judgment process. The focus of JA is wider, including the cue integration strategies that result in judgment. The following section will review the theoretical foundations and applications of JA as a research paradigm to study human judgment.

2.3.1. Brunswik's probabilistic functionalism and the lens model

In the 1940s and 1950s, psychologist Egon Brunswik proposed a sweeping reform to the direction of psychological research [41]. Contemporary psychologists were readily adopting analysis of variance as a statistical methodology to conduct systematic and controlled factorial designs to study human behavior while disentangling task and environmental variables. However, Brunswik viewed this trend as a means to describe the average organism, performing unrealistic tasks, under atypical laboratory conditions that were not representative of the organism's natural environment. Thus, results could be generalized to other average organisms not participating in the study, but could conclude little about other conditions or contexts for the behavior in question [40], [41].

Brunswik's novel approach away from mainstream psychology was coined *probabilistic functionalism*. His primary focus was to investigate the *relationships* between the organism and its environment, which is based on probabilistic relations among environmental variables (i.e., cues). Brunswik theorized that cues in the environment could never be perfect indicators of a distal state of the environment (i.e., criterion). Thus the probabilistic relationship between cues and criterion was termed ecological validity, which indicates the potential utility of cues (i.e., for an organism to use). Similarly, the cues are also only probabilistically related to the organism's response and the application of cues was termed functional validity. In Brunswik's view, the degree of success (or achievement) indicates the extent of which an organism's utilization of the cues matches the ecological validity of the cues.

Another key aspect to Brunswik's theory is the presence of intra-ecological correlations among cues, indicating environmental redundancy. The presence of these correlations implies that if an organism integrates the cues differently on successive occasions, that there can still be a high degree of achievement permitted by trade-offs in emphasis among correlated cues. Brunswik referred to this concept as vicarious functioning [40].

Brunswik created the lens model as a tool for representing and summarizing the concepts involved in probabilistic functioning. The model provides symmetric descriptions of the environmental criterion and the organism, as both are related to cues (Figure 3).

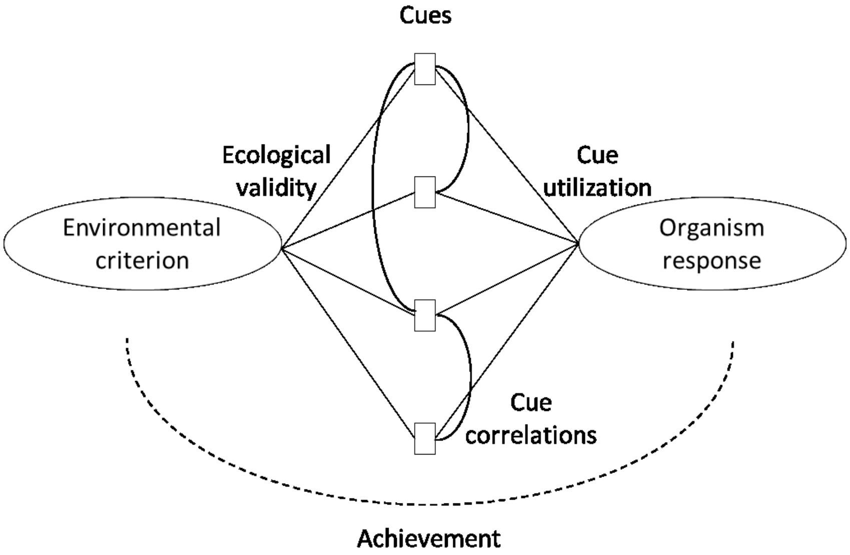


Figure 3. Brunswik's lens model.

The relationships between the cues and the criterion (ecological validities) and the cues and the organism’s response (cue utilizations) are depicted. There may be correlations among the cues and a subset of these is shown. Achievement is indicated by the correlation between the organism’s response and the environmental criterion. Again, this is maximized when the cue utilizations match ecological validities.

A methodological consequence of Brunswik’s theory was the use of *representative design*. This requires that both representative samples of the environment (cues and criterion) and representative samples of the organism be used in experimental designs. Statistical analysis must then support inferences with respect to situations in the ecology and to organisms.

To analyze the relationships between the organism and its environment, Brunswik advocated for an idiographic-statistical approach. He maintained that "individuals should be examined and statistically tested before attempting to generalize behavioral trends" [42] across organisms. Although Brunswik never specified a precise statistical methodology to accomplish this idiographic-statistical analysis, it is reported that he made indirect references to the possibility of multiple regression methods in his 1956 book [40], [41], [43].

Brunswik limited the application of his theories to human perception; however, his work has influenced research in a variety of areas, most notably human judgment. Kenneth Hammond was the psychologist primarily responsible for applying Brunswik's concepts to investigate human judgment.

2.3.2. Hammond's Social Judgment Theory

Hammond first applied Brunswik's theories to investigate clinical judgment [44]. He employed the lens model representation to summarize correlational results of a study investigating clinical judgments of patient IQs based on four cues (characterizing the patients). These judgments were compared to a criterion derived from statistical predictions using the Wechsler-Bellevue (W-B) IQ test scores. This first application demonstrated the utility of the lens model to explore three aspects of human judgment: understanding how the clinicians' judgments correlated with the W-B scores (the criterion), exploring the use of multiple regression to quantify how well clinicians could combine cue information to make a clinical judgment, and identifying cue correlations indicating the possibility of vicarious functioning.

This work laid the foundation for expanding the application of the lens model to other paradigms, including multiple-cue probability learning (MCPL) [45], interpersonal conflict (IPC) [46] and interpersonal learning (IPL) [47]. The latter two paradigms gave rise to the idea that Brunswik's theories could be applied to the social domain of human judgment, and thus *Social Judgment Theory* (SJT) was formed [48]. This opened the door to use Brunswik's lens model to investigate and compare multiple judges, even when no environmental criterion is available (e.g., correspondence between judges, differences in cue utilizations, etc.).

Hammond's work in SJT particularly drew attention to the possible forms of relationships between the cues and the criterion, and the cues and the human judgment. He showed how the functional form and the cue weights could be separated using multiple regression procedures[2] [48]. Around the same time, Goldberg [50] also described how linear models sufficiently predicted clinical judgments even when the human judges reported using more complex, non-linear strategies. Simple linear models have also out-performed judgment strategies based on logical rules [51], even when human judges verbalize their strategies in terms of logical rules and insist they are using a more complicated strategy than a linear additive model would suggest. Linear models have also replicated process-tracing models of judgment that were thought to be more cognitively representative [52] and even non-optimal linear models have accurately predicted human judgments[53].

Human judgment research with a Brunswikian focus using regression tools has been applied in many domains, including education [54–56], health care [57–59], accounting [60], [61], risk [62], social welfare [63], meteorological forecasting [64], and public project evaluations [65]. Because not all of these applications investigated the social aspect of judgment (between judges), the research paradigm became more generally known as *judgment analysis (JA)*.

[2] Bottenberg and Christal [49] are generally credited with the first use of multiple regression to analyze human judgment [40].

2.3.3. Judgment analysis

Judgment analysis (JA), as a modern research paradigm, is particularly well suited for investigating the relationships between cues, the human judge, and the environmental criterion. It is also appropriate for judgments that fall in the quasi-rational region of the cognitive continuum theory of judgment [42]. These judgments are defined by possessing some analytical features that are defensible and some intuitive features that are not completely traceable. Judgments of this nature could include judging the quality of hypertension care based on patient population measures that possess both analytical (e.g., average blood pressure reading) and intuitive features (e.g., expecting that many patients are not compliant with recommended medications).

JA can be characterized by the design of the system under investigation and represented with different versions of the lens model. These include the single, double, triple, and n-systems designs.

2.3.3.1. Single system lens model

The single system design constitutes a methodology for investigating human judgment when no environmental criterion is available [40]. It is represented using the lens model framework as shown below where the criterion (E) is grayed out.

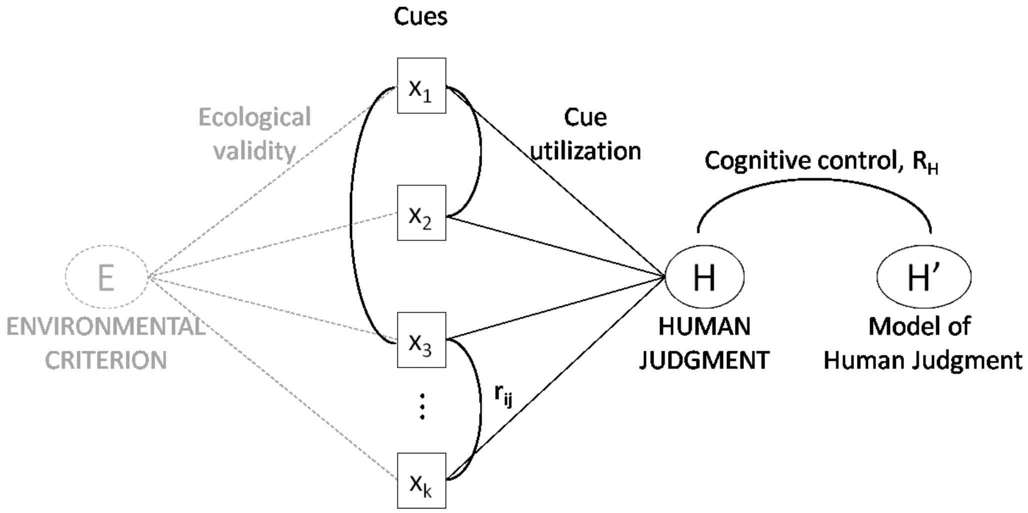


Figure 4. Single system lens model.

Cooksey [40] defines three valid reasons why a criterion may not be available. First, it may be impossible to define or measure a criterion. This may occur under simulated judgment cases, cases representing future judgment situations that are not currently performed, or cases regarding future environmental states where it is unrealistic to wait for the criterion to be measured. Second, there may be confidentiality, ethical, or legal reasons for the unavailability of the criterion. For example, Brady and Rappoport [66] studied judgments related to nuclear fuel theft potential from nuclear power plants. It would have been unsafe to use actual criterion measures of theft potential in their study in the event of misuse or theft of their data. Last, criterion measures may not be incorporated into judgment studies in circumstances where the criterion is irrelevant to the research goals. For example, research that is conducted explicitly to examine the cognitive systems of judges may not incorporate criterion measures. Under these circumstances, the environmental predictability may not be characterized.

Although judgment achievement cannot be measured under these circumstances, at the individual judge (idiographic) level, one can still investigate different aspects of human judgment (H) after a human has judged a sample of cue profiles. These aspects include cue utilization (e.g., the linear model relating cues to judgments), the weighting applied to each cue to make a judgment, and the degree of cognitive control the judge uses in applying his or her judgment strategy (R_H). Using linear regression techniques, R_H is the coefficient of multiple correlation of the regression on the human's judgments with the cue values, or the correlation of judgments and predictions of judgments based on the model. Cue correlations (r_{ij}) in the environment can also be investigated.

There are occasions when idiographic analysis is insufficient to answer research objectives. For example, if specific experimental groups of judges have been constructed (i.e., when treatment groups have been constructed to systematically vary an independent variable of interest), then idiographic analysis is only the first step. Nomothetic comparisons between the various idiographic measures may be performed to investigate the impact of different treatment groups. This is typically accomplished using univariate or multivariate analysis of variance (ANOVA) techniques [40].

2.3.3.2. Double system lens model

The double system design constitutes the classic Brunswik representation. This common form of the lens model provides symmetric descriptions of the environmental criterion and the human judge as both are related to cues (Figure 5).

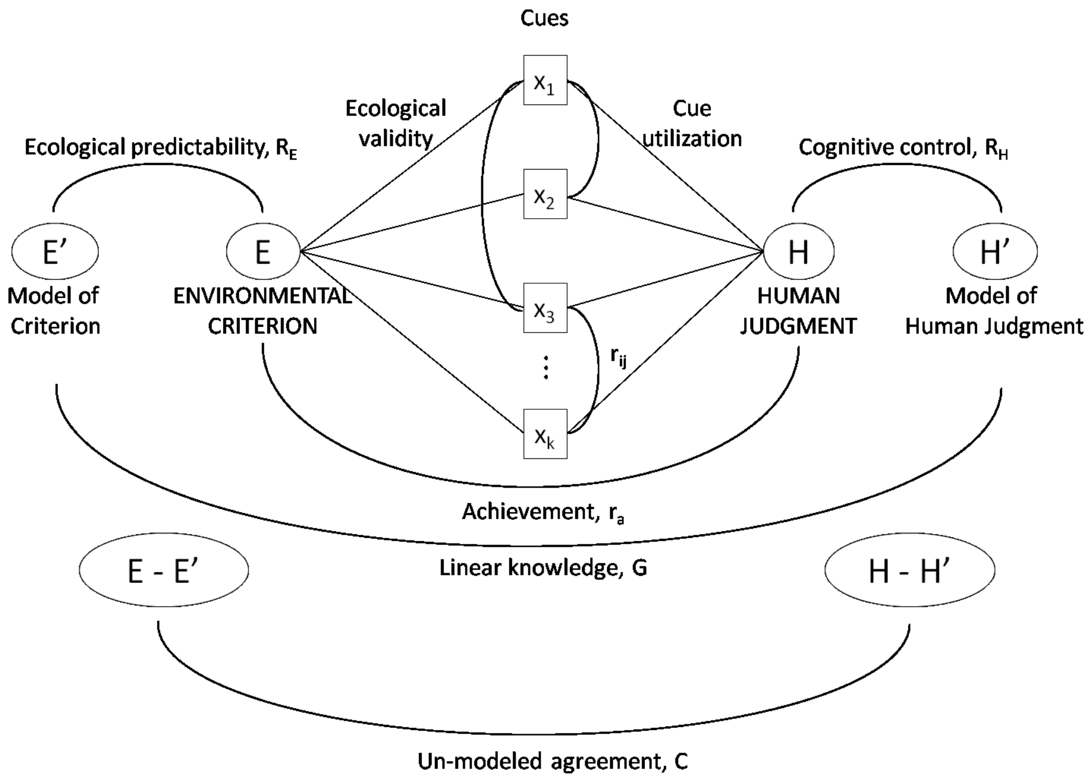


Figure 5. Double system lens model.

Similar to the single system design, the double system allows one to investigate the cue utilization of the judge, the weighting applied to each cue, the cognitive control (R_H), and cue correlations (r_{ij}). With the criterion available, one can similarly investigate the ecological validity of the cues (function relating cues to criterion) and the ecological predictability (R_E). Using linear regression techniques, R_E is the coefficient of multiple correlation of the regression on the criterion with the cue values, or the correlation of criterion values and predictions of the criterion based on the environmental model.

The double system lens model also provides a framework to calculate other judgment performance measures. Correlations representing achievement (r_a), linear knowledge (G) and un-modeled agreement (C) can also be determined. Achievement is the correlation between the human's judgment and the environmental criterion and is maximized when the cue utilizations match ecological validities. Linear knowledge (G) is the correlation between predictions from the criterion model and the human judgment model. This represents the extent of a judge's understanding of the task properties with the cues in the model. Un-modeled agreement (C) is the correlation between the residuals of the two models. A non-zero value of *C* could indicate common reliance on cues not included in the model, non-linearity of cue function forms, common cue interactions, or chance agreement between random model errors for both the criterion and the human judgment.

Both idiographic and nomothetic analyses can be performed with a double system design as mentioned in the single system design section above.

2.3.3.3. Triple and n-system lens model

The triple system lens model (Figure 6) is typically used to investigate the social aspects of human judgment. In this design there are two independent judges, which could be human or automated judges. All measures possible in the single and double system design are also possible in this design for each independent judge. Further, comparisons can be made between the judges [40]. The agreement between judgments made by Judge A and Judge B is indicated in the figure. This lens model representation has been primarily used in studies to investigate interpersonal conflict (IPC) [46], interpersonal learning (IPL) [47], and human-automation interaction [7], [11].

The n-systems lens model is a logical extension to the triple system to account for more than two independent judges.

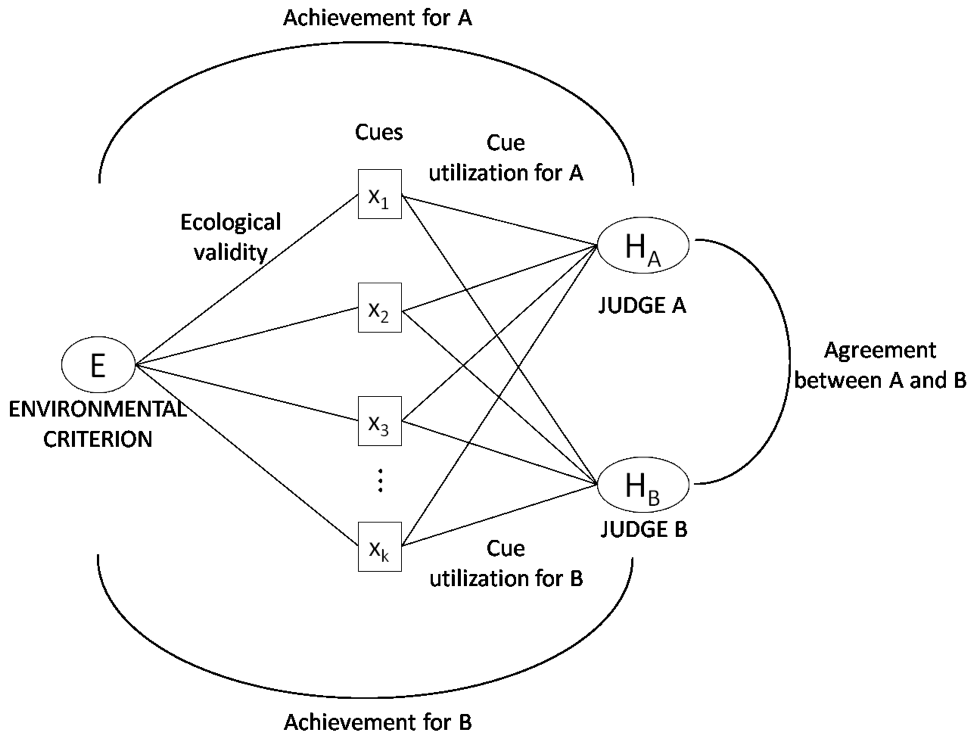


Figure 6. Triple system lens model.

2.3.4. The lens model equation

The lens model with its associated statistical parameters provides a conceptual and a methodological tool to study human judgment in many contexts. Taking advantage of regression statistics derived from the lens model concepts, the Lens Model Equation (LME) was developed by Hursch, Hammond, & Hursch [67] and Tucker [68]. It provides a mathematical basis for partitioning judgment achievement into the lower-level correspondences that account for the contributions of the environment and of the human judge. The LME assumes that the human and the criterion have been modeled with multiple linear regression using the cues as inputs to both models.

The LME (equation 2.1) separates achievement into two multiplicative components: the linear component that represents achievement attributed to the linear modeling of the judge and the ecology and the configural component that represents un-modeled aspects of achievement (i.e., the unpredictability of both the criterion and the human judge and the degree of which un-modeled aspects of the two models correspond). The variables in the equation are the correlations between components of the lens model as depicted in Figure 5. Again, because achievement is a correlation, the highest possible value is one.

$$r_a = GR_E R_H + C\sqrt{1 - R_{\mathrm{E}}^2}\sqrt{1 - R_{\mathrm{H}}^2} \tag{2.1}$$

If there is little un-modeled agreement (C approximately equal to zero), the LME can be simplified to the product of three correlations: linear knowledge (G), ecological predictability (R_E), and cognitive control (R_H).

$$r_a \approx GR_{\mathrm{E}}R_{\mathrm{H}} \tag{2.2}$$

If achievement is less than one, it can then be decomposed to understand why judgment performance is not perfect. For example, a judge may demonstrate an achievement of 0.80 on judging the identity of an approaching aircraft. As equation 2.2 above indicates, it could be that the judge's knowledge of the environment is limited (G = 0.80) but the environment is fully predictable (R_E = 1.0) and the judge is perfectly consistent in executing his or her judgment strategy (R_H = 1.0) based on the cues. On the other hand, the judge might have perfect linear knowledge and cognitive control but be working in an environment that is less than fully predictable given the cues (R_E = 0.80). Finally, it could be the case that the environment is fully predictable and the judge has perfect knowledge but makes judgments inconsistently (R_H = 0.80). Typically it is a combination of these effects in actual judgment analysis studies.

Table 3 summarizes the components of the lens model that can be derived and investigated within the judgment analysis framework and the lens model representation.

Table 3. Lens Model Equation parameters.

Component	**Name**	**Description**
r_a	Achievement	Correlation between the human's judgments and environmental criterion
G	Linear knowledge	Correlation between predictions from the models of the human judge and the environmental criterion
R_E	Ecological Predictability	Coefficient of multiple correlation of the regression on the criterion with the cue values
R_H	Cognitive Control	Coefficient of multiple correlation of the regression on the human's judgments with the cue values
C	Un-modeled Agreement	Correlation between the residuals (or errors) of the criterion model and the human judgment model

2.3.5. The expanded lens model (ELM)

In an attempt to better understand and investigate forecasting judgments, Stewart and Lusk [69] expanded the basic lens model framework by inserting two additional sets of cues (Figure 7). First, they inserted “true descriptors” between the cues and the criterion. Second, they inserted “subjective cues” between the cues and the human judgment.

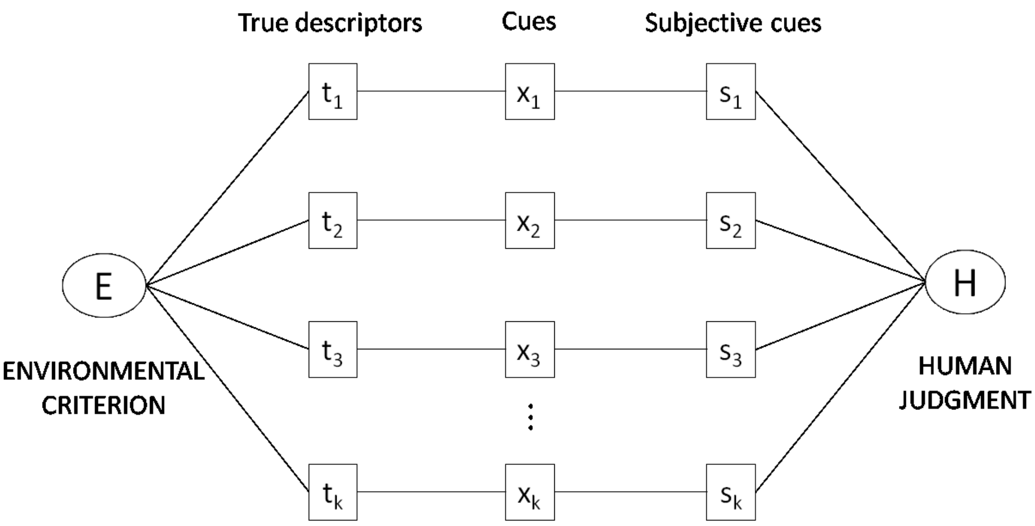


Figure 7. Expanded lens model (ELM).

The “true descriptors” are meant to model the true cues in the environment, possibly different from those available to the judge. For example, the true speed of an approaching air craft might be 608 mph (the “true descriptor” or true cue), but because radars are imperfect in sampling the atmosphere and processing data, the cue might be 591 mph. This expansion acknowledges that true descriptors are not always directly available to the judge as cues. Using this model, the environmental criterion can be correlated with predictions from either a model of the criterion using the true descriptors (R_{Et}) or a model of the criterion using the cues (R_{Ex}).

The “subjective cues” are meant to model subjective interpretation of the cues, which may be different from the cues. For example, the approaching air craft speed on a display might be 591 mph, yet a person may read the display as 600 mph. This expansion acknowledges that different human judges may process cues differently based on their subjective interpretation. Similar to above, they point out that the human’s judgments can be correlated with either predictions from a model of judgment using subjective cues (R_{Hs}) or a model of judgment using the cues (R_{Hx}).

Based on the LME, we know that judgment achievement (correlation between environmental criterion and human judgment) will be degraded if environmental predictability or cognitive control is low. With the expanded lens model, Stewart and Lusk [69] suggest that low environmental predictability (traditionally characterized by R_{Ex}) could be due to either an inadequate relationship between the environmental criterion and the true descriptors or between the true descriptors and the cues. Thus, they decompose R_{Ex} into the product of R_{Et} and the ratio of R_{Et} to R_{Ex}. They term this ratio "fidelity of the information system," or V_{tx}, to characterize the relationship between true descriptors and cues.

$$R_{Ex} = R_{Et}\frac{R_{Ex}}{R_{Et}} = R_{Et}V_{tx} \tag{2.3}$$

Similarly, low cognitive control (traditionally characterized by R_{Hx}) could be due to either inadequate relationships between the cues and subjective cues or between subjective cues and the human judgment [69]. Thus, they also show that R_{Hx} can be decomposed into the product of R_{Hs} and the ratio of R_{Hx} to R_{Hs}. They term this ratio "reliability of information acquisition," or V_{sx}), to characterize the relationship between cues and subjective cues.

$$R_{Hx} = R_{Hs}\frac{R_{Hx}}{R_{Hs}} = R_{Hx}V_{sx} \tag{2.4}$$

This expansion to the lens model has conceptual and theoretical value. However, the ELM does not explicitly consider the role of information automation to support the human in their judgment task. Also, to our knowledge, the ELM has only been empirically tested in two contexts [10], [9], [70]. In the first application [10], [9], the authors used a reduced version of the ELM (i.e., no true descriptors). They modeled human judgment under two conditions and found that the (reduced) ELM was useful in indicating differences in judgment performance between the two conditions due to environmental predictability that would not have been uncovered with using only the standard judgment performance measures of achievement or skill score [71].

In the second application [70], the authors replaced the criterion and the true descriptors with another human judge and another set of subjective cues (i.e., making the model symmetric, with the cues in the center). They found this application useful to identify differences in how individual judges used the cues (poor correlation between judge models or low value of G).

2.4. Judgment analysis as a method to inform automation design

Judgment analysis and the lens model conceptualization have been used specifically to support automation design under limited contexts. Primarily, judgment analysis has been used as a framework to investigate IA systems acting as independent judges. It has also been used to compare IA judgment strategies to those of human judges and to subsequently examine the interaction between human judge and automated judge [7], [8], [11–13].

For example, HAJL (Human Automated Judge Learning) is one quantitative, methodological framework that builds directly on the triple lens system design from judgment analysis and the concepts from interpersonal learning (IPL) [7]. Its goal is to study human interaction with an automated judge, rather than another human judge as was traditionally investigated in IPL experiments. Specifically, HAJL includes three experimental phases and measures to capture relevant features of the human's judgment processes in conjunction with an automated judge in order to inform IA design.

In the training phase, HAJL focuses on modeling the human judge with no interaction with the automated judge. In the interactive learning (IL) phase, the human and automation first make independent judgments. The human can then view the automation's judgment and make a subsequent "joint" judgment. The correlation between their independent judgments measures their (lack of) conflict. The correlation between the human's independent and joint judgments provides a measure of (lack of) compromise and the correlation between the automation's judgment and the joint judgment measures the extent to which the human adapts to the automation's judgment (Figure 8). This phase supports quantitatively measuring use of an IA judgment through the compromise and adaptation measures, which is different from others who have used subjective techniques to measure automation use [72].

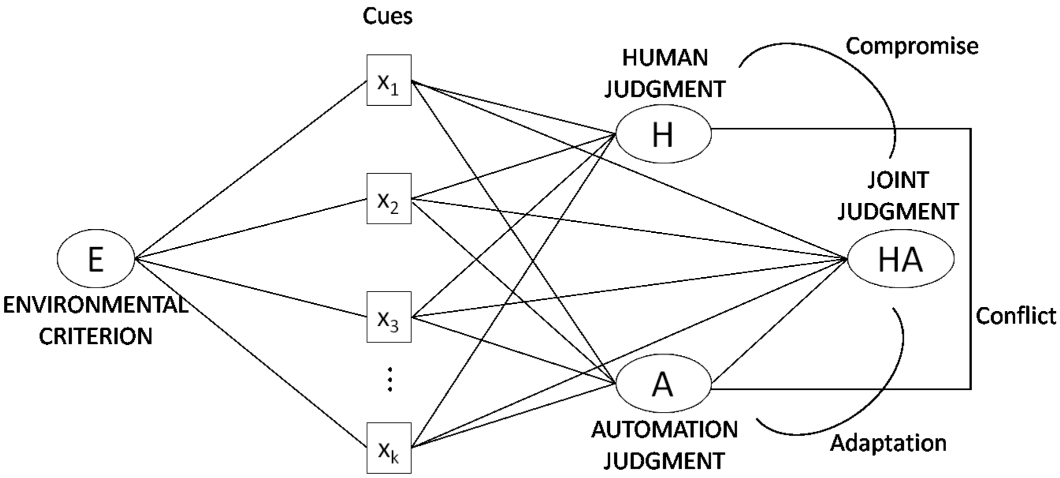


Figure 8. Lens model used in IL phase of HAJL.

In the prediction phase, the human has no interaction with the automated judge. The human makes an independent judgment and a judgment of what the automated judge would have predicted had it been available. This phase supports quantitatively measuring the human's understanding of the automation through measures of predictive accuracy (correlation between the automation's judgment and the human's prediction of it) and similarity (correlation between the human's judgment and the human's prediction of the automation or the automation's actual judgment).

HAJL measures in combination can help one understand the human-IA judge interaction [7]. For example, patterns in the unaided judgments and compromise/adaptation measures can identify over-, under-, or appropriate reliance on the IA judge. In one experiment employing HAJL, unaided judgment achievement in the IL phase was much lower than the automation's judgment achievement. This may indicate an appropriate context for the human to adapt their judgment to an IA judgment.

The HAJL framework has only been employed to investigate human-automation interaction where the human and IA act as independent judges [7], [73]. While useful in such contexts, it forces a certain experimental procedure where IA supports the human judge in only one phase and the IA always provides an algorithmically-generated judgment for the human to consider. IA could also be designed to support lower-level cognitive processes involved in the judgment process, such as gathering cues, presenting cues in various ways, and/or providing interpretation of cues. Further, HAJL does not take into consideration the different cue sets in the expanded lens model suggested by Stewart and Lusk [69] to account for discrepancy between true descriptors, available cues, and subjective cues.

2.5. Empirical findings of human interaction with information automation

Several studies have investigated the effects of automation " type" (as defined in [1]) on human performance [2–5], [74–76]. Favorable results have been found for using information automation compared to decision automation, particularly with imperfect automation [77–80]. However, very few studies have investigated judgment performance when different cognitive judgment functions are supported by IA (i.e., varying the level of IA support or the content of information presented to the human judge).

IA could be designed to support perception of cues using graphical or numerical displays. This level of IA has obvious benefits over not providing IA for tasks where it is impractical for an unaided human to perceive cues, such as in health care, aviation, and process control [81–83]. IA designed to support perception of cues using highlighting of a sub-set of cues has led to both improved judgment achievement [84–86] and degraded judgment performance in terms of regression bias (i.e., inappropriate narrowing of the judgment range) [8] and slower response times when the IA failed to highlight compared to conditions that never used IA cue highlighting [87].

An approach to supporting comprehension of cues has been advocated by a growing group of researchers in human-automation system design [81], [82], [88–90]. The notion here is that instead of having the automation provide a judgment, it will instead aid the human in understanding the important relationships in the task environment. However, this ecological approach has mainly been applied in the design of complex process and supervisory control systems and not with IA to support human judgment.

There may also be benefits from IA that provides an automated assessment [7], [8], [91]. Bass and Pritchett [7] found that judgment performance improved for an air traffic conflict prediction task when humans were provided with an automated assessment compared to when they were not provided with an automation assessment. This result was replicated by Horrey et al. [8] in a threat assessment task. However, they also found that half of the participants overused the IA support by not recognizing that the IA provided a poor judgment on one trial. Similar acceptance of an incorrect automated judgment was also observed by Layton et al. [91].

Judgment performance may also benefit from IA providing information explaining how it integrates cues to derive its automated assessment. Seong and Bisantz [12], [13] found that even IA that does not consistently apply its judgment strategy and is not consistently accurate can enhance human judgment performance over unreliable IA that does not explain its strategy in air traffic identification tasks. However, Bass and Pritchett [7] did not replicate this result of finding any added benefit when explanation of IA strategy was provided to support air traffic conflict judgments. Conversely, they observed that participants who were not provided with an explanation of IA strategy could better predict the IA's judgments. The same group of participants also had higher judgment achievement when no longer provided with an IA judgment compared to other participants who had previously been provided with both an IA judgment and explanation of the IA strategy.

Thus, while some studies suggest that providing an IA judgment and insight into the IA's judgment strategy may improve human judgment performance, there may also be situations where supporting perception and comprehension of the cues in the environment without providing a judgment improves human judgment performance. We also must consider that the usefulness of increasing information may stop at a limited amount and that additional information may actually hinder the judgment process [92], [93]. Thus, tradeoffs in what functionality the IA employs (and what information it displays to the human) is critical.

2.6. Summary of conceptual gaps in the literature and research aims

Given that IA systems to support human judgment will continue to proliferate and we know that the design of automated systems impacts human behavior (often in unanticipated ways), we need frameworks to understand human judgment supported by IA to inform automation design choices.

However, the approach of using frameworks of types and levels of automation to guide design and evaluation are only useful as a starting point because they have focused on automation to support human decision making and supervisory tasks. No taxonomies exist to describe how the level of automation to support judgment may vary. Parasuraman et al. [1] even point out that the development of a taxonomy to describe levels of IA support is a major challenge for the human factors community. Additionally, while the evaluation measures suggested by such frameworks are important for human-automation systems in general, these measures are often subjective (e.g., workload, situation awareness, and trust) and do not reflect human judgment performance as comprehensively as described in the judgment analysis paradigm where judgment achievement can be decomposed into other quantitative measures that reflect consistency, task knowledge, and environmental predictability.

Further, the approach of using judgment analysis and the lens model to characterize human judgment does not explicitly consider IA that aids the human in the cognitive processes involved in judgment. Rather, the focus has been with automation acting as an independent judge. No quantitative model has yet been developed to investigate IA supporting different cognitive functions of human judgment (e.g., perceiving cues, comprehending cues, and integrating cues to form a judgment). Further, no model has accounted for discrepancies between how true cues (in the environment) are transformed to available cues (used by IA) to displayed cues (presented by the IA and used by the human judge) to subjective cues (interpreted by the human judge).

Last, there is little empirical work (none conclusive) explicitly comparing the effects on human judgment performance under varying conditions of IA support. Some studies suggest that providing an IA judgment and insight into the IA's judgment strategy may improve human judgment performance. However, there may also be situations where supporting perception and comprehension of the cues in the environment improves judgment performance.

Thus, this research has three aims. The first aim is to develop a new framework for the design and evaluation of information automation to support judgment, titled the Expanded Lens Model with Automation (ELMA). To provide design support, ELMA accounts for discrepancies between how cues in the environment are transformed into displays to operators via automated processes. The transformation is based upon the desired, hierarchical level of cognitive support (i.e., from cue perception, to comprehension, to assessment). In addition to design support, ELMA also includes quantitative measures to evaluate the human-automation system with an idiographic-statistical approach. Multiple linear regression and correlation analysis are employed to characterize achievement, consistency, and knowledge of the human judge; potential and accuracy of the automation; and predictability of the environment. Nomothetic analysis can then be applied to investigate the effect of automation design choices on general judgment performance.

The second aim is to demonstrate ELMA's utility and address domain-specific objectives, by evaluating an existing system that supports air traffic controllers in *judging the probability of an air traffic conflict* using heading and speed cues. Specifically, the objectives were to investigate the effect of providing additional levels of support, in tandem with an automated judgment, on achievement and consistency.

The third aim is to demonstrate ELMA's utility to understand and inform the design of automation to support physicians in *judging the quality of population-based hypertension care* using cues on patient outcomes (e.g., blood pressure) and processes (e.g., medications prescribed). Specifically, the objectives of this aim were: 1) to identify cues needed to judge hypertension care quality, 2) to understand how these cues differentially influence judgment, and 3) to evaluate the effect of level support on judgment achievement, consistency within individual physicians, and reliability across physicians.

Results will provide further empirical data regarding the impact of automated support on judgment performance, inform the design of IA for multiple domains, and ultimately aid in enhanced human-automation judgment performance.

3. Introducing the Expanded Lens Model with Automation (ELMA)

3.1. Context for using ELMA

In some settings unaided human judgment is inadequate or even impossible. For example, it would be unrealistic for an unaided physician to make a judgment regarding the proportion of patients up to date on mammography screenings in a large clinic. To make such a judgment, the physician might need to access each individual patient's medical record, mentally note the date of her last mammogram, determine if this date was within the recommended guidelines for frequency of screening, and then subsequently integrate all patient records to arrive at a total proportion of patients up to date. Thus, utilizing automation to acquire and/or integrate cues is beneficial (or even necessary) in contexts where unaided human judgment is unrealistic.

However, no automated system can be designed to be perfectly reliable under all circumstances. For example, there are many sources of uncertainty in cues used to make health care quality judgments. Cues may be missing (e.g., due to having a mammograms at another facility), erroneous (e.g., due to inaccurate input to medical records), or confounded by variables outside of the physicians' control (e.g., mammograms were recommended, but patients failed to go to appointments because of lack of financial resources). Compounding these issues, if the automation is thought to be reliable by the human user, the user may misuse or inappropriately trust the automation [94].

The ELMA framework and methodology is developed with these circumstances in mind - where the automation supports the human judgment process in some way, but the human judge makes the final judgment. Most often these judgments can be defined as quasi-rational in that they possess some analytical features that are defensible and some intuitive features that are not completely traceable [40], [42].

When designing automation to support the human judge under such contexts, one important system design question to ensure accurate and coherent judgments is:

> *What level of support should the automation provide the human in the judgment process (e.g., should the automation only acquire and present information, or should it integrate information and present an algorithmically-generated judgment for the human to accept or veto)?*

Conceptual frameworks for defining levels of automated judgment support and quantitative methodologies for analyzing human-automation systems are needed to answer these questions and guide the system design and evaluation process. Because no commonly accepted frameworks or systematic procedures have been developed to respond to this need, ELMA has been developed to fill this gap. The effectiveness of the joint human-automation judgment system is determined by the interplay of the automation, the human judge, and the environmental context. Thus, ELMA provides:

1. A representation of the judgment task environment
2. A representation of the human's judgment process
3. A representation of the automation's role in supporting the human judge
4. A means to conceptualize an experimental design space from where one can derive and test hypotheses regarding automation design choices related to the level of automated support (in terms of information presented on the human-automation interface)
5. A procedure to guide the design of automation to support human judgment
6. Quantitative measures to evaluate the human-automation system

ELMA is a useful tool for systems engineers because the framework unifies the sensor designer, the algorithm designer, the display designer and the judgment trainers responsible for the complete design and evaluation of human-automation judgment systems.

3.2. ELMA model to conceptualize human judgment supported by automation

The ELMA model to guide automation design and evaluation expands on the judgment analysis paradigm [40] and the lens model [41] and expanded lens model representations [69] of human judgment. A generic representation of the ELMA extension to the expanded lens model is shown in Figure 9. “Cues” from Stewart and Lusk’s expanded model [69] have been decomposed into three components interposed between the “true cues” and the “subjective cues”: the “available cues,” “information automation” and the “displayed cues.”

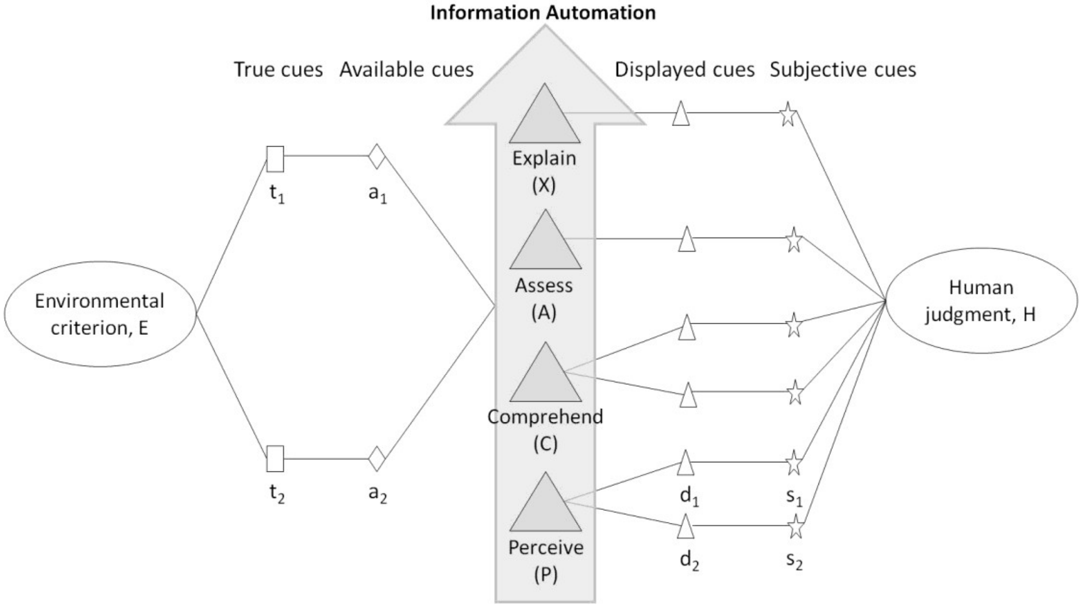


Figure 9. ELMA extension to the lens model.

The ELMA lens model depicts lines that represent the relationships between the environmental criterion, different types of cues, information automation, and the human judgment. Recall that the environmental criterion on the far left of the figure is defined as the true state of the environment that the human is attempting to judge. In a perfect setting, the human’s judgments on the far right of the figure (supported by IA) would always match the environmental criterion. This correlation between environmental criteria (**E**) and human judgments (**H**) is defined as the achievement of the human-automation system.

$$r_a = cor(\boldsymbol{E}, \boldsymbol{H}) \tag{3.1}$$

In a perfect setting the human-IA achievement would equal one. However, if this is not the case, we can assume that the breakdown in performance is due to one of the transformations from environmental criterion to human judgment, how the human uses the cues to arrive at a judgment, or how predictable the criterion is based on the cues. The next sections will discuss these transformations and contributions to judgment achievement.

3.2.1. Environmental criterion, true cues, and available cues

Starting at the left side of Figure 9, we define the true cues to be directly related to the environmental criterion, E, as

$$E = M_{E.t}(t_1, t_2, t_3, \ldots, t_n) \tag{3.2}$$

where t_i are the true cues and $M_{E.t}$ is a linear model that describes the relationship between the true cues and the criterion. In this definition, it is assumed that $M_{E.t}$ captures all of the relationships between the true cues and the criterion. The correlation between actual criterion values and predictions from the model of the criterion using the true cues represents the *true environmental predictability*. Theoretically, we would expect this correlation to be equal to one.

$$R_{E.t} = cor(\boldsymbol{E}, \boldsymbol{M}_{E.t}) \tag{3.3}$$

However, not all true cues will be readily available in the environment and there will be imperfect relations between the true cues and cues that are actually available, the available cues. Thus, equations 3.2 and 3.3 are more useful conceptually than analytically and the available cues can be expressed as a function of the true cues where *n* is the number of true cues.

$$a_i = f(t_i) \text{ for } \mathrm{i} = 1, \ldots, \mathrm{n} \tag{3.4}$$

Consequently, there is a probabilistic relationship between the available cues and the criterion, E, represented by

$$E = M_{E.a}(a_1, a_2, a_3, \dots, a_n) + E_{E.a} \tag{3.5}$$

where a_i are the available cues, $M_{E.a}$ is a linear model that describes the relationship between the available cues and the criterion, and $E_{E.a}$ represents the residuals of the model. The correlation of the actual criterion values with predictions from model of the criterion using the available cues is defined as the *available environmental predictability* and is useful for characterizing the context in which the automation system is able to support the human's judgments.

$$R_{E.a} = cor(\boldsymbol{E}, \boldsymbol{M}_{E.a}) \tag{3.6}$$

The ratio of $R_{E.a}$ to $R_{E.t}$ will theoretically be between 0.0 and 1.0 because the available cues will never be better predictors of the environmental criterion than the true cues. Stewart and Lusk define this measure as the fidelity in the measurement system [69], but in our case we define this ratio as the *automation potential* (ranging from 0 to 1) in that it represents the potential of the automation to support the human judge through the accurate transformation of true cues to available cues. If we assume that the true cues can be used to perfectly model the criterion, then this ratio is reduced to $R_{E.a}$.

$$V_{a.t} = \frac{R_{E.a}}{R_{E.t}} \tag{3.7}$$

The measure of automation potential, $V_{a.t}$, characterizes the transformation between true cues and available cues, which could be addressed with sensor design or data acquisition techniques.

3.2.2. A taxonomy of automated judgment support

Moving to the right of the available cues in Figure 9, the information automation (IA), depicted by the arrow and triangles, uses the available cues to support the human judgment process. The level of IA support varies in terms of how the available cues are transformed into displayed cues that the human judge can use. To discuss, design, and analyze IA employing different levels of support, it is useful to have a taxonomy. A taxonomy characterizes the automation's role in supporting the human judge. For ELMA, this is represented through four levels of judgment support: *perceive, comprehend, assess, and explain*. These levels represent a means to systematically examine the effect of IA judgment support on human judgment performance. Further, we also suggest that the levels need not be implemented exclusively and that combinations of levels may be used to support human judgment.

The "perceive" level of IA support aids the human judge in perceiving the available cue(s) from the environment, or in directly transforming the available cue(s) (a_i) to displayed cue(s) (d_i) for the human judge to consider when making a judgment. At the "perceive" level, the displayed cues can be expressed as a function of the available cues where n is the number of available cues.

$$d_i = f(a_i) \text{ for i} = 1, \ldots, \text{n} \tag{3.8}$$

The transformation from available to displayed cue(s) could be a direct transformation in that the available cue(s) are displayed to the precision that they are available to the automation, using a particular display representation. However, the automation could also modify the available cue value in some way (e.g., round to the nearest whole number). With more than one available cue, the automation could also choose to present the displayed cues simultaneously or with a particular temporal sequencing. Automation support at this level is particularly beneficial for tasks where it is not feasible for the human to gather available cues on their own and this has been applied in domains such as health care, aviation, and process control [81–83].

The "comprehend" level of IA support goes beyond simple perception of available cue(s) and supports comprehending meaning by comparing the current available cue(s) with either a pre-set standard or threshold value (c) or to descriptive information about the available cue(s) under consideration. For example, support at this level could include presenting the relationship between an available cue ($a_{i,\,t}$) and the value of available cues one time epoch prior ($a_{i,\,t\text{-}1}$). Automation support at this level also supports working memory (i.e., it requires working memory to compare a cue value to a threshold value or to previous value of the cue under consideration) and also provides context-dependent information to the human judge [95].

$$d_i = f(a_i, c) \text{ or } d_{i,t} = f\left(a_{i,t,}, a_{i,t-1}\right) \text{ for i} = 1, \ldots, \text{n} \tag{3.9}$$

The “assess” level of IA support aids the human judge by combining available cues to make an automated assessment or judgment (defined as A) of the environmental criterion based on the available cues and presents this to the human judge. Such automated judgment systems have been found to be valuable, particularly as alerting systems [96], [97].

$$d_i = f(a_1, a_2, \ldots, a_n) \text{ for } i = 1, \ldots, n \tag{3.10}$$

If an IA system provides this level of judgment support, we can specifically characterize the automation’s judgment performance by calculating its achievement (correspondence with the environmental criterion).

$$r_{aA} = cor(\boldsymbol{E}, \boldsymbol{A}) \tag{3.11}$$

We can also model the automation’s judgment using multiple linear regression with the available cues as inputs to the model (equation 3.12). Using this model, we can characterize the automaton’s cognitive control (equation 3.13) by correlating the automation’s judgments with predictions from the model of the automation’s judgments. Also, if we model the environmental criterion using available cues, we can also measure the automation’s linear knowledge (equation 3.14) and un-modeled agreement (equation 3.15).

These measures are useful to characterize judgment performance specifically of the automation at this level of support. Although the performance of the automation may affect the performance of the human judge, we subsequently focus more on the human’s judgment performance to characterize the human-automation system, as it is the human who makes the final judgment.

$$A = M_{A.\text{a}}(a_1, a_2, a_3, \ldots, a_n) + E_{A.\text{a}} \tag{3.12}$$

$$R_{A.a} = cor(\boldsymbol{A}, \boldsymbol{M}_{A.a}) \tag{3.13}$$

$$G_{A.a} = cor(\boldsymbol{M}_{E.\mathrm{a}}, \boldsymbol{M}_{A.a}) \tag{3.14}$$

$$C_{A.a} = cor(\boldsymbol{E}_{E.a}, \boldsymbol{E}_{A.a}) \tag{3.15}$$

The "explain" level of judgment support provides the highest level of automated support to the human judge by presenting the cue combination strategy of how the automation arrived at the judgment of the environmental criterion, or the specific relationship between available cues (a_i) and automated judgment (A). This could be in the form of the linear regression weights for each available cue that the automation uses in arriving at its judgment, with a particular display representation. This level of automation support has been used in air traffic control and air traffic identification judgment tasks (e.g., [12], [13], [96], [98–101]). This level of support could also include comparing the automated judgment with either a pre-set standard or threshold value(c) or to descriptive information about the automated judgment.

$$\begin{gathered} d_i = g\big(f(a_1, a_2, \dots, a_n)\big) \\ d_i = g(f(a_1, a_2, \dots, a_n), c) \\ d_i = g(f(a_1, a_2, \dots, a_n)_t, f(a_1, a_2, \dots, a_n)_{t-1}) \end{gathered} \tag{3.16}$$

In some judgment contexts, the IA will support the human at least at the "perceive" level. There could be contexts where the human is supported solely by higher levels, particularly the "assess" level. However, for this dissertation we focus on tasks where the human works *with* the IA to arrive at a judgment, which will likely include also having access to the available cues (through the IA "perceive" level). Table 4 summarizes the taxonomy of information automation support to human judgment as defined by four levels.

Table 4. Taxonomy of information automation (IA) support to human judgment.

Level of information automation (IA) judgment support	**Functionality of IA**	**Functional description of displayed cues presented by IA**
PERCEIVE (P)	• Display current value(s) of available cues simultaneously • Display current value(s) of available cues with temporal sequencing	• $d_i = f(a_i)$ for $\mathrm{i} = 1, \dots, \mathrm{n}$ (simultaneously) • $d_i = f(a_i)$ for $\mathrm{i} = 1, \dots, \mathrm{n}$ (sequentially)
COMPREHEND (C)	• Display the comparison of an available cue(s) to a pre-set standard value(s) • Display the comparison of an available cue(s) to a descriptive statistic(s) of itself	• $d_i = f(a_i, c)$ for $i = 1, \dots, n$ • $d_i = f\left(a_{i,t,}, a_{i,t-1}\right)$ for $\mathrm{i} = 1, \dots, \mathrm{n}$
ASSESS (A)	• Display an automated judgment as a result of combining current available cue values	• $d_i = f(a_1, a_2, \dots, a_n)$ for $i = 1, \dots, n$
EXPLAIN (X)	• Display available cue combination strategy • Display the comparison of the automated judgment value to pre-set standard value • Display the comparison of an automated judgment value to descriptive statistics of itself	• $d_i = g\left(f(a_1, a_2, \dots, a_n)\right)$ for $i = 1, \dots, n$ • $d_i = g(f(a_1, a_2, \dots, a_n), c)$ for $i = 1, \dots, n$ • $d_i = g(f(a_1, a_2, \dots, a_n)_t, f(a_1, a_2, \dots, a_n)_{t-1}$ for $i = 1, \dots, n$

3.2.3. Displayed cues, subjective cues, and the human judgment

After transforming the available cues through one or more levels of judgment support, the IA then presents <u>displayed cues</u> to the human judge. The displayed cues are the specific representations of the available cues as described above in Table 4. We can model the environmental criterion using the displayed cues as follows.

$$E = M_{E.d}(d_1, d_2, d_3, \dots, d_n) + E_{E.d} \tag{3.17}$$

Above, $M_{E.d}$ is a linear model that describes the relationship between the displayed cues and the criterion, and $E_{E.d}$ represents the residuals of the model. The correlation of criterion values with predictions from the model of the criterion using the displayed cues can be defined as the *displayed environmental predictability* and is useful for characterizing the context in which the human makes judgments.

$$R_{E.d} = cor(\boldsymbol{E}, \boldsymbol{M}_{E.d}) \tag{3.18}$$

Recall from equation 3.3 that $R_{E.t}$ is the true environmental predictability. Thus, we know that the ratio of $R_{E.d}$ to $R_{E.t}$ will theoretically be between 0.0 and 1.0 because the displayed cues will never be better predictors of the environmental criterion compared to the true cues. We define this ratio as the *true accuracy of the automation*.

$$V_{d.t} = \frac{R_{E.d}}{R_{E.t}} \tag{3.19}$$

Because the true cues may not be available (and equation 3.19 not useful), we can also define the ratio of $R_{E.d}$ to $R_{E.a}$ as the *displayed accuracy of the automation* (equation 3.20). This also ranges from 0 to 1 because we can assume that the displayed cues will never be better predictors of the environmental criterion compared to the available cues. The displayed accuracy of the automation characterizes the transformation from available cues to displayed cues. We might expect this relationship to be close to 1. However, there could be circumstances where this is not the case, such as if the algorithms used to transform available cues to displayed cues are extracting data at a rate less than the update rate of the available cues. This could be addressed with algorithms to process the available cues into displays for the human judge.

$$V_{d.a} = \frac{R_{E.d}}{R_{E.a}} \tag{3.20}$$

Similar to our model of the environmental criterion, we can model the human's judgments using the displayed cues. This describes the correspondence between the human's judgments and the displayed cues.

$$H = M_{H.d}(d_1, d_2, d_3, \dots, d_n) + E_{H.d} \tag{3.21}$$

$M_{H.d}$ is a linear model that describes the relationship between the displayed cues and the human's judgment and $E_{H.d}$ represents the residuals of the model. The correlation of the human's judgments with predictions from the model of the human judgments using the displayed cues can be defined as the *displayed cognitive control* and is useful for characterizing the consistency with which the human judge employs his or her judgment strategy based on the displayed cues.

$$R_{H.d} = cor(\boldsymbol{H}, \boldsymbol{M}_{H.d}) \tag{3.22}$$

The human judge will see the displayed cues, yet may interpret these in different ways, due to factors such as displayed cue representation, prior experience, bias, motivation, or training. The subjective cues are meant to represent this subjective interpretation and inherent transformation of the displayed cues and can be expressed as a function of the displayed cues.

$$s_i = f(d_i) \text{ for i} = 1, \dots, \text{n} \tag{3.23}$$

We can also model the human's judgments using the subjective cues where $M_{H.s}$ is a linear model that describes the relationship between the subjective cues and the human's judgment and $E_{H.s}$ represents the residuals of the model.

$$H = M_{H.s}(s_1, s_2, s_3, \dots, s_n) + E_{H.s} \tag{3.24}$$

Gathering the subjective cue values would require asking the human judge to report their perception of each displayed cue. To our knowledge, this has been only been done in two experiments using judgment analysis [9], [70] as it is a cumbersome and unintuitive task for the human judge. Further, the above definition of subjective cues (equation 3.23) assumes that all subjective cues are a function of displayed cues. However, there may also be subjective cues that the human uses that are not provided by the automation (i.e., not via displayed cues). If known, these additional subjective cues can also be incorporated into the model. This may be useful for both informing the design of the automation (i.e., update to include the additional cues, if possible) or to inform training (i.e., investigating how the operator is using the additional subjective cues).

We define the correlation of the human's judgments with predictions from the model of the human judgments using the subjective cues as the *subjective cognitive control.*

$$R_{H.s} = cor(\boldsymbol{H}, \boldsymbol{M}_{H.s}) \tag{3.25}$$

Similar to how we characterized the accuracy of the automation, we can also characterize the accuracy of the human judge in transforming displayed cues to subjective cues. The ratio of displayed to subjective cognitive control ($R_{H.d}$ to $R_{H.s}$) is defined as the *displayed accuracy of the human judge.* Theoretically, this ratio will be between 0.0 and 1.0 because the displayed cues will never be better predictors of the human's judgment compared to the subjective cues.

$$V_{d.s} = \frac{R_{H.d}}{R_{H.s}} \tag{3.26}$$

The transformation between displayed cues and subjective cues could be addressed with display representations, training, or motivation for the human judge.

We can also model the environmental criterion using the subjective cues, as follows where $M_{E.s}$ is a linear model that describes the relationship between the subjective cues and the criterion and $E_{E.s}$ represents the residuals of the model.

$$E = M_{E.s}(s_1, s_2, s_3, \dots, s_n) + E_{E.s} \tag{3.27}$$

The correlation of criterion values with predictions from the model of the criterion using the subjective cues can be defined as the *subjective environmental predictability* and is useful for characterizing the predictability of the environment based on the human judge's subjective interpretation of the environment.

$$R_{E.\mathrm{s}} = cor(\boldsymbol{E}, \boldsymbol{M}_{E.\mathrm{s}}) \qquad (3.28)$$

3.2.4. The lens model equation using displayed cues

With the ELMA extension to the expanded lens model representation, we can consider two other correlations to characterize the human-automation system. These correlations use displayed cues because we are particularly interested in investigating the impact of IA support, manifested as displayed cues, on judgment performance in order to inform IA design choices. However, it should be noted that one could also measure the following correlations with other cue sets (i.e., true, available, or subjective cues).

The first correlation is between predictions from the environmental criterion model using the displayed cues and the human judgment model using the displayed cues (equations 3.17 and 3.21). We define this as the *displayed linear knowledge* ($G_{H.d}$) and it represents the extent to which the criterion values and judgment values would agree if both models using the displayed cues were perfect (i.e., $R_{E.d}$ = $R_{H.d}$ = 1). This can also be thought of as the extent that a judge understands the judgment task ecology based on the automation's level of support (through displayed cues) to ensure correspondence of their judgment strategy with the displayed predictability in the environment.

$$G_{H.d} = cor(\boldsymbol{M}_{E.d}, \boldsymbol{M}_{H.d}) \qquad (3.29)$$

The second additional correlation is the *displayed un-modeled agreement*, $C_{H.d}$. This is the correlation between the residuals of the two models using displayed cues. A non-zero value of $C_{H.d}$ could indicate common reliance on cues not included in either model, non-linearity of cue function forms, common displayed cue interactions, or chance agreement between random model errors for both the criterion and the human judgment.

$$C_{H.d} = cor(\boldsymbol{E}_{E.d}, \boldsymbol{E}_{H.d}) \tag{3.30}$$

In some judgment contexts, the human factors engineer will only have access to the displayed cues. If this is the case, we can still make use of the lens model equation (LME). Recall that achievement is defined as the correlation between environmental criterion (E) and human judgment (H). This can then be decomposed using the correlations defined in the previous section involving the displayed cues.

$$r_a = G_{H.d}R_{E.d}R_{H.d} + C_{H.d}\sqrt{1 - R_{E.d}^2}\sqrt{1 - R_{H.d}^2} \tag{3.31}$$

If the displayed un-modeled agreement ($C_{H.d}$) is low, the LME can be simplified to the product of three correlations: displayed linear knowledge ($G_{H.d}$), displayed ecological predictability ($R_{E.d}$), and displayed cognitive control ($R_{H.d}$).

$$r_a \approx G_{H.d}R_{E.d}R_{H.d} \tag{3.32}$$

This decomposition provides insight in understanding why judgment performance is not perfect. For example, a judge may demonstrate an achievement of 0.80 on judging the proportion of patients up to date on mammography screenings. As equation 3.32 above indicates, it could be that the judge's knowledge of the environment based on the displayed cues is limited ($G_{H.d}$ = 0.80) but the environment is fully predictable with the displayed cues ($R_{E.d}$ = 1.0) and there is perfect transformation between true cues, to available cues, to displayed cues. Also, the judge is perfectly consistent in executing his or her judgment strategy based on the displayed cues ($R_{H.d}$ = 1.0) and there is perfect transformation between displayed and subjective cues.

On the other hand, the judge might have perfect displayed linear knowledge and displayed cognitive control but be making a judgment based on displayed cues that are not fully predictable of the criterion ($R_{E.d}$ = 0.80). This could be due to an inadequate relationship between the true cues and the environmental criterion (characterized by $R_{E.d}$) or to a weak transformation between either true and available cues (characterized by $V_{a.t}$) or available and displayed cues (characterized by $V_{d.a}$). Thus, we can decompose the displayed environmental predictability into these three components as follows.

$$R_{E.d} = R_{E.t}V_{a.t}V_{d.a} = R_{E.t}\frac{R_{E.a}}{R_{E.t}}\frac{R_{E.d}}{R_{E.a}} \tag{3.33}$$

Finally, it could be the case that the environment is fully predictable with the displayed cues and the judge has perfect displayed linear knowledge but makes judgments inconsistently ($R_{H.d}$ = 0.80). This could be due to an inadequate relationship between the subjective cues and the human's judgment (characterized by $R_{H.s}$) or to a weak transformation between the displayed cues and the subjective cues (characterized by $V_{d.s}$). Thus, we can decompose the displayed cognitive control into these components as follows.

$$R_{H.d} = R_{H.s}V_{d.s} = R_{H.s}\frac{R_{H.d}}{R_{H.s}} \tag{3.34}$$

We can include these decompositions, based on the ELMA framework, into the LME to obtain the following equation to understand the lower level contributions to human-IA judgment achievement.

$$r_a \approx G_{H.d}R_{E.t}\frac{R_{E.a}}{R_{E.t}}\frac{R_{E.d}}{R_{E.a}}\frac{R_{H.d}}{R_{H.s}}R_{H.s} \approx G_{H.d}R_{E.t}V_{a.t}V_{d.a}V_{d.s}R_{H.s} \tag{3.35}$$

3.2.5. Summary of ELMA measures

Table 5 summarizes the quantitative measures of the ELMA expansion to the extended lens model that can be derived and investigated within the ELMA framework. While all measures are useful to characterize the environment, the automation, and the human judge, the table also includes a column to indicate various implications for specific IA design choices. As mentioned previously, the ELMA framework and associated measures provide structure to investigate human judgment supported by automation. ELMA also unifies the sensor designer, the algorithm designer, the display designer and the judgment trainers responsible for the complete design of human-automation judgment systems.

Table 5. ELMA parameters.

	ELMA measure	Description	Cue set needed	Specific implication for automation design
r_a	human-IA judgment achievement	correlation between environmental criteria (E) and human judgments (H)	none	
$R_{E.t}$	true environmental predictability	correlation between environmental criteria (E) and predictions from the model of the criterion based on true cues ($M_{E.t}$)	true	
$R_{E.a}$	available environmental predictability	correlation between environmental criteria (E) and predictions from the model of the criterion based on available cues ($M_{E.a}$)	available	data acquisition or sensor design
$V_{a.t}$	automation potential	ratio of $R_{E.a}$ to $R_{E.t}$	true and available	data acquisition or sensor design
$R_{E.d}$	displayed environmental predictability	correlation between environmental criteria (E) and predictions from the model of the criterion based on displayed cues ($M_{E.d}$)	displayed	algorithms to process available cues

$V_{d.t}$	true accuracy of the automation	ratio of $R_{E.d}$ to $R_{E.t}$	true and displayed	algorithms to process available cues
$V_{d.a}$	displayed accuracy of the automation	ratio of $R_{E.d}$ to $R_{E.a}$	available and displayed	algorithms to process available cues
$R_{H.d}$	displayed cognitive control	correlation between human judgments (H) and predictions from the model of the human judgment based on displayed cues ($M_{H.d}$)	displayed	level of IA support, display representations, training to improve consistency of judgments
$R_{H.s}$	subjective cognitive control	correlation between human judgments (H) and predictions from the model of the human judgment based on subjective cues ($M_{H.s}$)	subjective	training to improve consistency of judgments
$R_{E.s}$	subjective environmental predictability	correlation between environmental criteria (E) and predictions from the model of the criterion based on subjective cues ($M_{E.s}$)	subjective	training to understand the judgment task context
$V_{d.s}$	subjective accuracy of the human judge	ratio of $R_{H.d}$ to $R_{H.s}$	displayed and subjective	display representations, training to interpret the displayed cues

$G_{H.d}$	displayed linear knowledge of human judge	correlation between predictions from the environmental criterion model and the human judgment model using the displayed cues	displayed	training to better understand the judgment task
$C_{H.d}$	displayed un-modeled agreement	correlation between the residuals of the environmental criterion model and the human judgment model using displayed cues	displayed	

Implications for IA design mentioned in the fourth column of Table 5 have been evaluated in the literature to some extent regarding their impact on human judgment performance. For example, better sensors or data acquisition techniques impact the available predictability of the environmental criterion and thus the automation potential measure. Decreased environmental predictability has shown to be associated with decreased consistency in judgment [102].

Specific display designs (in terms of both the content and representation of information) have been shown to impact judgment achievement in numerous applications (e.g., [19], [91], [103]). More generally, the amount of information on displays has shown to impact judgment performance in that judgment performance may actually deteriorate beyond a certain amount of information [92], [93]. Training, particularly using cognitive feedback methods [104] has also been shown to be beneficial to improving judgment performance [105–107].

However, few studies have explicitly investigated the impact that the level of IA support has on human judgment [7], [8], [13]. Thus, this research is particularly interested in this aspect of automation design.

3.3. ELMA method to design and evaluate information automation

The ELMA framework of human judgment is combined with a series of steps and an iterative procedure to guide the design and evaluation of information automation. This component of ELMA extends the work of Parasuraman et al. [1] as an empirical method to address the system design question of choosing the level of IA judgment support to provide while also maintaining Brunswik's requirements for a representative and systematic methodology in order to generalize results to the judgment context. The ELMA methodology (in the context of the PSW framework for types of automation, Figure 2) is presented in Figure 10. Step 3 of the method is not the focus of this dissertation and is therefore grayed out in the figure.

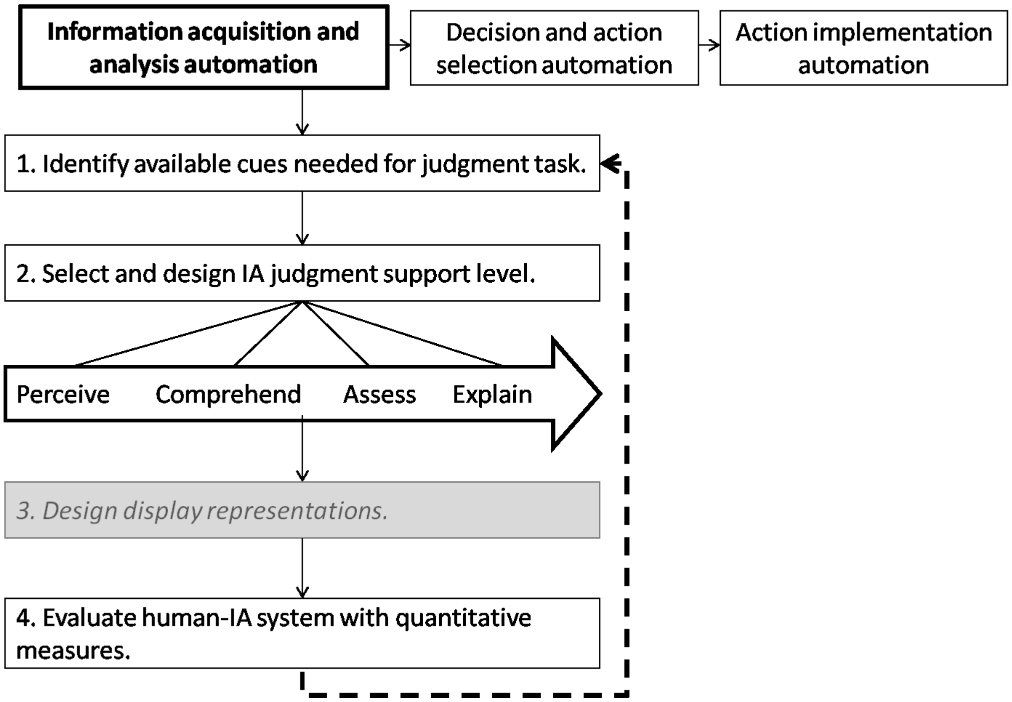

Information acquisition and analysis automation
Decision and action selection automation
Action implementation automation
1. Identify available cues needed for judgment task.
2. Select and design IA judgment support level.
Perceive
Comprehend
Assess
Explain
3. Design display representations.
4. Evaluate human-IA system with quantitative measures.

Figure 10. ELMA method to guide design and evaluation of information automation.

The first step of the ELMA method involves identifying the available cues pertinent to the judgment task. Cooksey [40] identified four methods to identify cues for a given judgment task that have been used in lens model-based experiments: interviews and surveys, document analysis, objective analysis of the ecology, and verbal protocol analysis. These methods are also suggested as viable methods to identify available cues as defined in the ELMA framework.

Given a set of available cues, the IA support level is then designed to transform the available cues into displayed cues for the human judge to use in his or her judgment process. As discussed in the previous sections, automation can support the human at one or more levels described in the taxonomy presented in Table 4. Because there is little to no empirical data suggesting what level of judgment support results in better human-automation judgment performance, it is difficult to suggest guidelines for this step. Further, there is probably no simple answer as there will likely be tradeoffs between costs and benefits, depending on the judgment context. Thus, the automation designer may select a level or combination of levels based on their understanding of the judgment task ecology. However, any particular combination of levels should be evaluated using the ELMA measures as discussed below. The best performing combination of levels could then be selecting for further refinement and evaluation.

The display representations are likely to be specific to the judgment task of interest as well. There have been many empirical studies suggesting different display principles based on task characteristics and we do not go into detail regarding these as this is not the focus of this dissertation (hence the grayed out step 3). For a review, see [26–28] or the artful books by Tufte [108], [109]. Display representations may be specifically evaluated with surveys, heuristic evaluations, and usability studies [110]. However, in the context of ELMA, we evaluate the human-automation system as encompassing both the content and representation of displayed cues, which allows us to test different displays.

The methodological consequences of this procedure include the ability to derive hypotheses regarding design choices, particularly related to the level of IA judgment support. To test such hypotheses or to evaluate the human-automation system as represented by the ELMA framework (step 4), a judgment analysis experiment is required. Cooksey [40] provides thorough guidelines for conducting such an experiment. We review only the most relevant aspects here and modify them to appropriately fit within the ELMA framework.

The first step in evaluating the human-automation system is to create judgment profiles. This involves obtaining sets of available cue values. For example, if we were evaluating the human-automation system that judges the probability that a patient has influenza, we would need to create multiple judgment profiles (or patient profiles) for this task. This could be profiles of multiple patients presenting with influenza symptoms (or cues) at a single clinic. Cooksey recommends a ratio of 10 judgment profiles to every one available cue needed for the judgment task due to the desire to use multiple linear regression to model both the criterion and the human judgment. Thus, if three available patient cues were needed to judge the likelihood that a patient has influenza (e.g., body temperature, presence of body ache, and presence of congestion), 30 patient profiles should be created.

The second step is to obtain criterion values for each judgment profile so that cue-criterion relationships may be modeled and achievement of the human-automation system (equation 3.1) can later be calculated. For our influenza example, the criterion values could be the actual diagnoses that the patients were given (e.g., influenza or not).

The third step is to instantiate the judgment profiles in the IA system or a prototype of the IA system (with the level of IA support under investigation). Some commercially available software is also available to conduct judgment analysis experiments (e.g., POLICY PC) [40]. Most importantly, an accurate method to collect the human judge's responses to the judgment profiles is necessary. Representative participants to participate must then be recruited. Time constraints and availability of participants (e.g., physicians) may create a challenge to maintaining an acceptable judgment profile number, depending on the average amount of time that is required for each judgment.

Once the judgment analysis experiment is completed, the ELMA measures presented in Table 5 can be calculated. As mentioned, true cues and subjective cues may not be readily available or convenient to obtain. However, insight into the human-automation system can still be gained using only the available and displayed cues.

Idiographic analysis may be conducted for specific human judges. This may involve analyzing the judgment performance of the best and worst judge (based on judgment achievement) in terms of how their cue weights, cognitive control, and linear knowledge vary. If a nomothetic evaluation is of interest (e.g., to investigate the effect of IA support on groups of participants), this can be conducted using standard analysis of variance techniques with the ELMA measures.

3.4. Conceptualizing a single cue judgment task example using the ELMA framework

To further explain the ELMA conceptualization of human judgment supported by automation, consider the example judgment task of assessing core body temperature. The IA used in this judgment task might be an oral digital thermometer. This task is represented with the ELMA lens model as shown in Figure 11. From left to right in the figure, the environmental criterion represents the actual state of the environment, or the actual core body temperature. For this task, there is one true cue that directly results from the environmental criterion, which is simply the true core temperature. However, this true cue may be different than what is available to the IA, represented by a transformation from true cue to available cue. In our example, the available temperature might be different than the true core temperature because we only have access to the oral temperature. This difference is depicted in the figure.

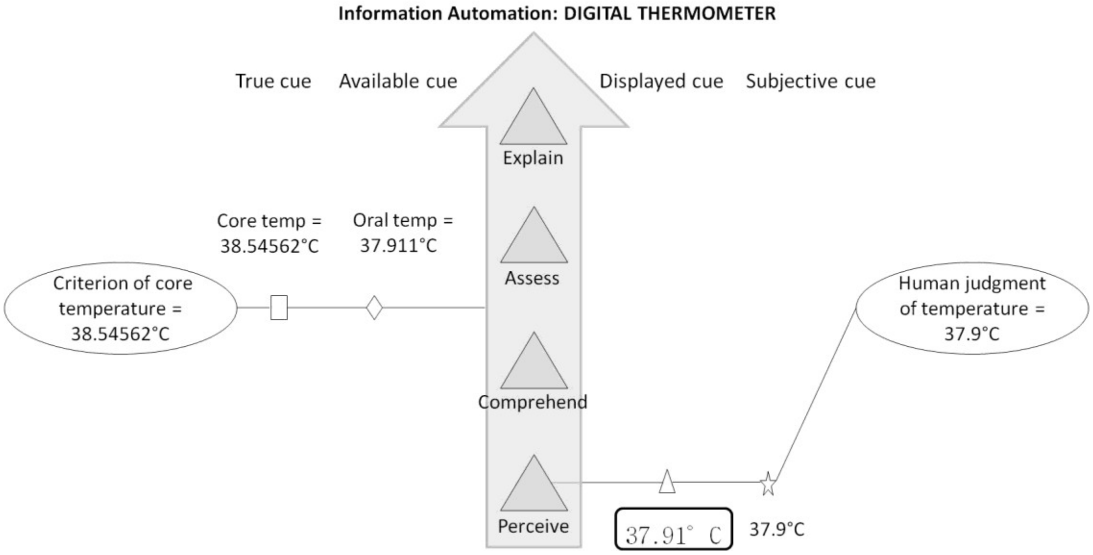


Figure 11. ELMA lens model representation of judgment task of assessing body temperature with IA providing “perceive” level of judgment support.

The IA (digital thermometer) can transform the available cue at the “perceive” level of judgment support, as shown in the figure. The IA then presents the available body temperature on a digital display for a human to perceive. In this example, there is a direct transformation from available to displayed cue. The human judge then subjectively interprets the cue (perhaps rounds to the nearest decimal place) and makes a judgment regarding body temperature.

The IA could further support the human judge in comprehending the cue value. This is depicted in Figure 12 by showing the available cue also passes through the “comprehend” level of judgment support. The digital thermometer compares the available body temperature to the known standard of normal, healthy body temperature and adds this to the display.

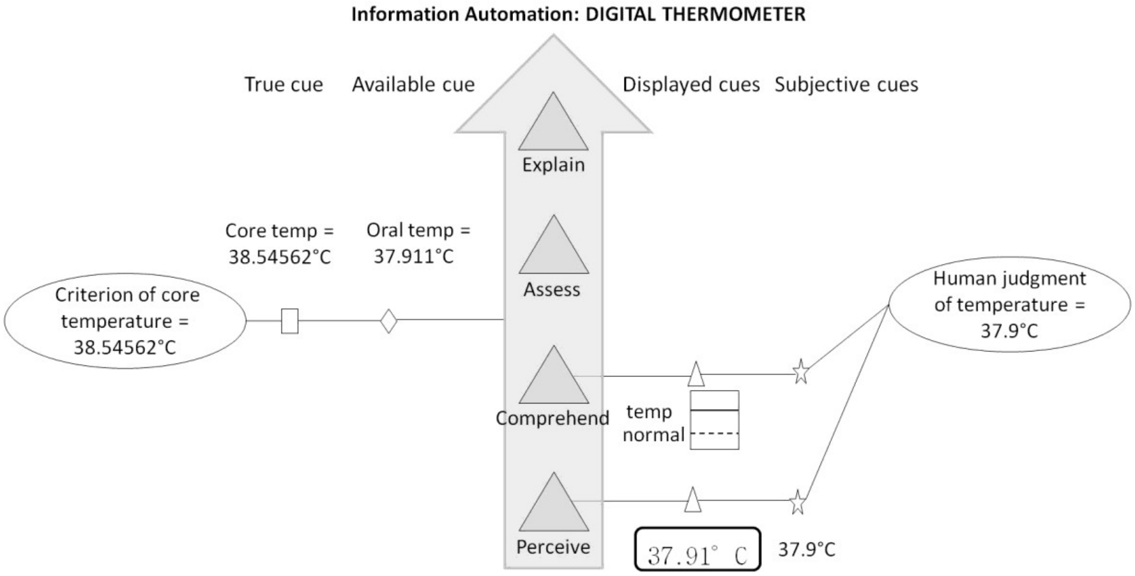


Figure 12. ELMA lens model representation of judgment task of assessing body temperature with IA providing “perceive” and “comprehend” levels of judgment support.

To evaluate this human-automation system, consider that a judgment analysis experiment was conducted. Because only one cue is required to judge body temperature, 10 judgment profiles were created, instantiated in a prototype thermometer, and presented to a single human to judge. For this example, we can imagine that we have access to every component of the ELMA framework. The data from this experiment are presented in Table 6.

Table 6. Profiles for a single cue judgment task example.

criterion	true cues	available	displayed	subjective	judgment
38.54562	38.54562	37.911	37.91	38.0	38.0
40.26941	40.26941	35.513	35.51	35.5	35.5
35.87664	35.87664	32.759	32.76	32.7	32.7
32.66984	32.66984	25.607	25.61	25.6	25.6
28.12587	28.12587	25.867	25.87	26.0	26.0
34.56541	34.56541	32.149	32.15	32.0	32.0
41.54213	41.54213	38.823	38.82	39.0	39.0
31.00254	31.00254	25.589	25.59	25.6	25.6
36.48369	36.48369	35.124	35.12	35.0	35.0
37.54562	37.54562	33.741	33.74	33.7	33.7

With these data, we can apply equations 3.1-3.35 to investigate human-automation system performance. First, we can correlate the human's judgments with the environmental criterion to obtain the human-automation judgment achievement.

$$r_a = 0.926 \tag{3.36}$$

Recall, that in a perfect setting, this result would be equal to one. Because our example does not have perfect human-automation judgment achievement we can investigate the components of the ELMA framework that may be influencing judgment achievement.

Our models of the environmental criterion and human judge are done with linear regression. We can model the environmental criterion using the true cues. Correlating the criterion with predictions from the model of the criterion gives us the true environmental predictability (or how predictable the criterion is based on true cues).

$$R_{E.t} = 1 \tag{3.37}$$

Similarly, we can model the criterion using the available cues and correlate the criterion with predictions of the criterion with the model using available cues to obtain the available environmental criterion.

$$R_{E.a} = 0.926 \tag{3.38}$$

The automation potential characterizes the transformation between true cues to available cues.

$$V_{a.t} = 0.926 \tag{3.39}$$

However, because $V_{a.t}$ is the ratio of two correlations, we may not notice any systematic differences in magnitude of the true cue set compared to the available cue set. If we look closely at the model of the criterion using available cues, we can see a magnitude shift in the model with an intercept, or constant regression parameter of ~10 (compared to zero with the true cues). Thus, the available cues are systematically lower than the true cues, which may be due to the sensors used to measures true cues. This shows value of modeling the criterion using the available cues.

We can also model the criterion using the displayed cues to obtain the displayed environmental predictability ($R_{E.d}$) and the true ($V_{d.t}$) and displayed accuracy ($V_{d.a}$) of the automation. Displayed accuracy of the automation characterizes the algorithms used to transform available to displayed cues.

$$R_{E.d} = 0.926 \tag{3.40}$$

$$V_{d.t} = 0.926 \tag{3.41}$$

$$V_{d.a} = 1 \tag{3.42}$$

We can model the human's judgments using the displayed cues and correlate the human's judgment to predictions of the human's judgment to obtain the displayed cognitive control of the human judge.

$$R_{H.d} = 1 \tag{3.43}$$

The human's judgments can also be modeled using the subjective cues to obtain the subjective cognitive control of the human judge and the subjective accuracy of the human judge. This characterizes the transformation from displayed to subjective cues.

$$R_{H.s} = 1 \tag{3.44}$$

$$V_{d.s} = 1 \tag{3.45}$$

To further investigate the human-automation performance for this task, we can also calculate the displayed linear knowledge ($G_{H.d}$) and un-modeled agreement ($C_{H.d}$).

$$G_{H.d} = 1 \tag{3.46}$$

$$C_{H.d} = -0.04 \tag{3.47}$$

We can interpret this to mean that the human judge has a perfect knowledge of the task (i.e., with only one cue, that is the cue used to make a judgment and thus, G is 1.0) and there is little un-modeled agreement between the criterion and human models.

In summary, using the lens model equation (as written in equation 3.35, which we can assume due to a value of $C_{H.d}$ close to zero), we know that judgment achievement is impacted by the transformation of true cues to available cues ($V_{a.t}$). The imperfect predictability of the criterion due to available cues could be due to imperfect measurement of core temperature using the oral thermometer and could be improved with a better sensing device.

3.5. Conceptualizing a multi-cue judgment task example using the ELMA framework

We can also consider a second judgment task example involving more than one cue to further explain the ELMA framework. A clinician may need to judge the number of daily calories that a patient in the intensive care unit needs. The IA used in this judgment task might be an electronic health record (EHR) with a calorie calculator module. This example is represented using the ELMA lens model as shown in Figure 13.

The environmental criterion is the actual daily calories needed, using indirect calorimetry (i.e., exact measurement of the number of calories needed at rest). In our example, this environmental criterion of 1529 is related to three true cues - exact measures of weight, height, and age of the patient. However, again, these exact measurements might not be available and there is a transformation from true to available cues. The IA then transforms these available cues through one or more levels of judgment support to present the human judge with displayed cues.

At the "perceive" level of judgment support (as shown in the figure) the IA supports the human in perception of the available cues by displaying the values of the available cues. With more than one cue available, the IA may also present the cue values in a specific temporal sequence in order to support perception.

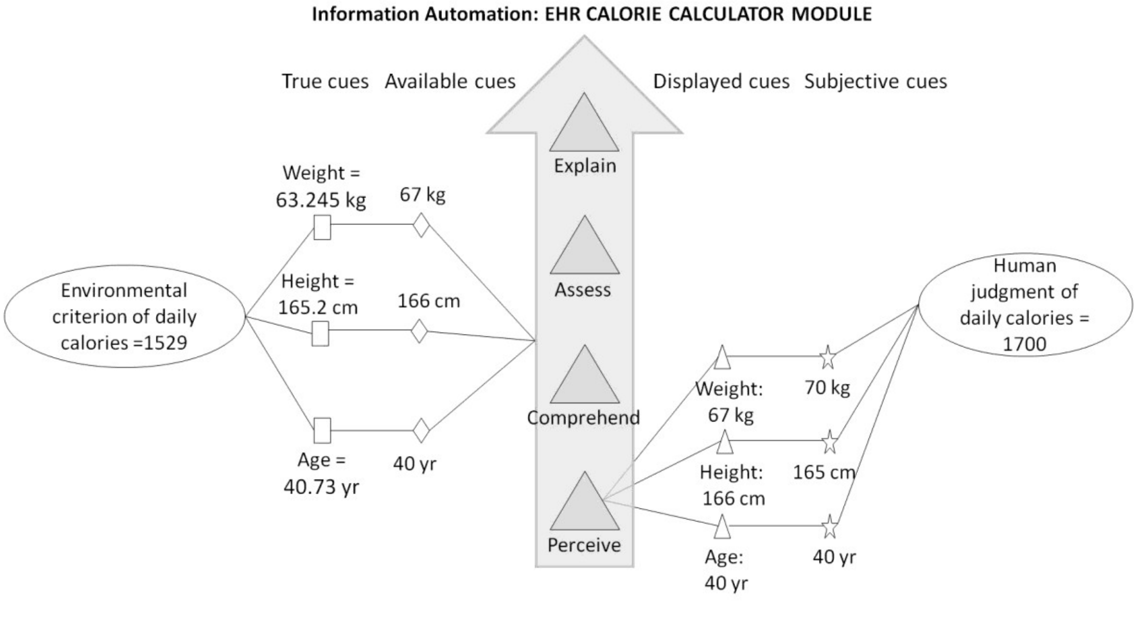


Figure 13. ELMA lens model representation of judgment task of assessing daily calories needed with IA providing "perceive" level of judgment support.

The IA may also support the human judge at the “comprehend” level of judgment support in our example as shown in Figure 14. At this level, the IA aids the human judge in comprehending the available cue through comparisons. These comparisons could include comparing the available cue values to pre-set standard values or to descriptive statistics of the available cues under consideration (such as average over the past year, as shown in the figure for two cues).

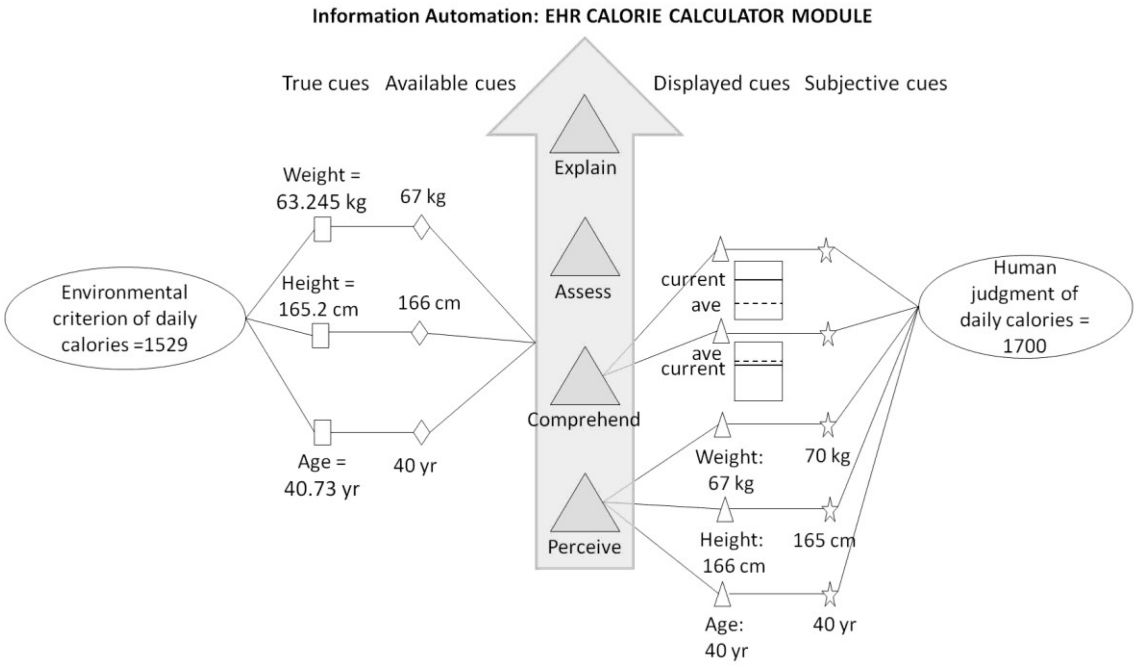


Figure 14. ELMA lens model representation of judgment task of assessing daily calories needed with IA providing “perceive” and “comprehend” levels of judgment support.

With more than one cue available, the IA may also support the human judge at both the “perceive” and “assess” levels of judgment support. This representation for our calorie assessment example is shown in Figure 15. At the “assess” level of judgment support, the IA aids the human judge in combining the available cue values to form automated assessment of the environmental criterion.

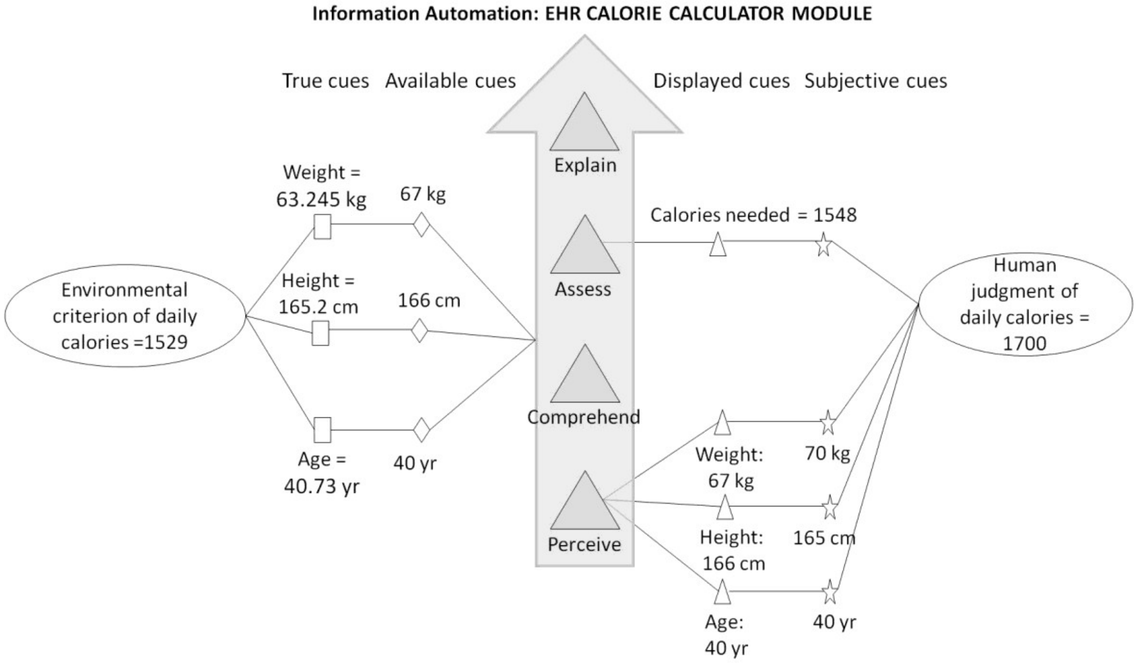


Figure 15. ELMA lens model representation of judgment task of assessing daily calories needed with IA providing “perceive” and “assess” levels of judgment support.

To evaluate this human-automation system, let us again consider that a judgment analysis experiment was conducted. Because three cues are required to judge calories needed, 30 judgment profiles were created, instantiated in a prototype IA system, and presented to a single human to judge. For this example, we can also imagine that we have access to every component of the ELMA framework. An excerpt of the data from this experiment is presented in Table 7.

Table 7. Profiles for a multi-cue judgment task example.

criterion	true	true	true	avail	avail	avail	disp	disp	disp	sub	sub	sub	judge	auto
	weight	height	age	weight	height	age	weight	height	age	weight	height	age		
1529	63.24	165.2	40.7	67	166	40	67	166	40	70	165	40	1700	1548
1413	45.78	150.8	34.6	41	149	34	41	149	34	40	150	34	1550	1146
1350	50.09	155.4	56.7	45	156	56	45	156	56	45	160	56	1300	1087
1708	90.87	180.8	43.8	93	177	43	93	177	43	90	180	43	1500	1940
1762	85.43	175.3	21.6	78	173	21	78	173	21	80	170	21	1740	1863
1542	75.08	164.2	47.1	81	165	47	81	165	47	80	165	47	1770	1688
1547	65.11	158.3	31.8	68	159	31	68	159	31	70	160	31	1750	1588
1604	82.58	179.3	56.7	88	176	56	88	176	56	90	180	56	1700	1779
1551	54.01	164.5	26.7	51	162	26	51	162	26	50	160	26	1700	1403
1329	50.99	161.1	67.2	52	164	67	52	164	67	50	160	67	1200	1149

We can correlate the human's judgments with the environmental criterion (shown below). Recall, that in a perfect setting, this result would be equal to one. Because our example does not have perfect human-automation judgment achievement we can investigate the transformation(s) that may be inaccurate.

$$r_a = 0.634 \tag{3.48}$$

Again, our models of the environmental criterion and human judge are done with multiple linear regression. We model the environmental criterion using the true cues. Correlating the criterion with predictions from the model of the criterion gives us the true environmental predictability (or how predictable the criterion is based on true cues).

$$R_{E.t} = 1 \tag{3.49}$$

Similarly, we can model the criterion using the available cues to obtain the available environmental criterion. In this example, $R_{E.a}$ is close to one. Thus, the automation potential, $V_{a.t,}$ is also close to one and this characterizes the transformation between true cues to available cues.

$$R_{E.a} = 0.99 \tag{3.50}$$

$$V_{a.t} = 0.99 \tag{3.51}$$

We can also model the criterion using the displayed cues to obtain the displayed environmental predictability. In this example, $R_{E.d}$ is also close to one, thus the true accuracy of the automation, $V_{d.t}$, is close to one and the displayed accuracy of the automation, $V_{d.a}$, is one.

$$R_{E.d} = 0.99 \tag{3.52}$$

$$V_{d.t} = 0.99 \tag{3.53}$$

$$V_{d.a} = 1 \tag{3.54}$$

We can also model the human's judgments using the displayed cues to obtain the displayed cognitive control of the human judge, $R_{H.d}$, as equal to 0.676. The human's judgments can also be modeled using the subjective cues to obtain the subjective cognitive control of the human judge, $R_{H.s}$, as equal 0.684. Thus, the subjective accuracy of the human judge, $V_{d.s}$ is equal to 0.988. This value characterizes the transformation from displayed to subjective cues.

$$R_{H.d} = 0.676 \tag{3.55}$$

$$R_{H.s} = 0.684 \tag{3.56}$$

$$V_{d.s} = 0.988 \tag{3.57}$$

To further investigate the human-automation performance for this task, we can calculate the displayed linear knowledge, $G_{H.d}$, and un-modeled agreement, $C_{H.d}$. These are 0.997 and -0.259 respectively. We can interpret this to mean that the human judge has a good knowledge of the task based on the displayed cues and the un-modeled agreement between the criterion and human models is fairly low.

Because the automation is also providing judgment support at the “assess” level, it is providing the human with an automated judgment of the criterion. We can also use lens model parameters to characterize the specific judgment performance of the automation (equations 3.11-3.15). In particular, we can calculate the automation’s achievement, cognitive control, and linear knowledge based on the available cues. These calculations show that in our example, the automation’s achievement, r_{aA} is equal to 0.937. Thus, the automation is providing the human with automated judgments well correlated to the criterion.

In summary, we know that the automation potential is nearly perfect. The displayed accuracy of the automation is also nearly perfect. Based on our calculations of displayed cognitive control, subjective cognitive control, and subjective accuracy of the human judge, we know that the transformation from displayed cues to subjective cues is the transformation contributing to less than perfect judgment achievement. We also know that the human judge is not perfectly consistent in applying their judgment strategy using their subjective cues. This inability of the human to consistently apply their judgment strategy could be due to imperfection in how the human interprets the displayed cues, or to other factors such as poor display representations, training, fatigue, use of other non-modeled cues such as physical exam findings, or bias.

3.6. Summary of the ELMA model and methodology

ELMA is a conceptual design and evaluation framework for automation that supports cognitive functions involved in judgment. It provides structure to understand and represent human-automation judgment systems and further enables the conceptualization of an experimental design space from where one can derive hypotheses regarding automation design decisions.

One advantage of ELMA is that it accounts for discrepancies between how cues in the environment are transformed into displays to operators via automated processes. The transformation is based upon the desired, hierarchical level of cognitive judgment support, from cue perception, to cue comprehension, to an automated assessment, to explanation of the automated assessment. This taxonomy complements the PSW framework for types and levels of automation to include levels specifically for information automation (IA), which have not yet been suggested. The idea that automation can be designed to support the human judge at different levels of support allows for a more explicit description about the role of the human and the role of the automation during the design and operation of human-automation judgment systems.

A second advantage is that ELMA includes quantitative measures to evaluate the human-automation system with an idiographic-statistical approach. Multiple linear regression and correlation analysis are employed to characterize achievement, consistency, and knowledge of the human judge, potential and accuracy of automation support, and predictability of the environment. Nomothetic analysis can then be applied to investigate the effect of automation design choices on general judgment performance. This also allows us to discuss quantitative, empirical judgment analysis-based results across different tasks and domains. More generally, our framework allows one to diagnose any positive or negative effects of different automation interventions on various objective measures of judgment performance.

ELMA makes several explicit claims about what features affect human-automation judgment performance:

1. Uncertainty in the judgment task (predictability of the criterion based on true cues, available cues, or displayed cues, and the potential the automation has to support the human judge)
2. Uncertainty in the ability of the automation to support the human judge (true and displayed accuracy of the automation)
3. Uncertainty in the strategy the human employs to combine cues to make judgments (displayed or subjective cognitive control and subjective accuracy of the human judge)

4. Uncertainty in the ability of the human to apply task knowledge (displayed linear knowledge)

We are particularly interested in the uncertainty related to the human judge (3 and 4 above). From this perspective, judgment performance could be influenced by the functionality of the automation (i.e., the level of IA support to aid in cognitive processes involved in judgment) or the design of the human-automation interface (i.e., displays). The former is of particular focus for this work.

The limitations of ELMA are that it requires the judgment task to be analyzed in accordance with the structure of the ELMA lens model (Figure 9). The judgment task must be defined as a known (or estimated) criterion to be judged, based on a set of cues. Further to implement the multiple regression procedure, judgments and criterion must be assigned quantitative values and numerical cues must be identified. The framework also makes high demands on the data that must be collected (i.e., 10:1 ratio of judgments to cues).

However, without the ELMA framework, analysis of judgment performance might only include measuring judgment accuracy or subjective appraisal of automated systems across groups of participants [1], [85], [96], [111]) or in comparing one design to another (e.g., [112], [113]). ELMA provides greater insight into human-automation system performance, such as how cues are used by individual judges, how consistently judgment policies are employed, and how well the human understands the task environment. This is important for the design of automated systems and to inform training interventions for human judges.

4. Applying ELMA to an air traffic control judgment task

This chapter discusses how the ELMA framework was applied to investigate an air traffic control judgment task to inform the design of IA to support air traffic control judgments. The ELMA model was used to characterize and analyze human-automation judgment performance under different conditions of IA judgment support. This involved re-analyzing data collected in a prior human subject experiment where participants made judgments about the probability of air traffic conflicts under different conditions of automated support [73].

4.1. Introduction

Air traffic controllers and pilots make judgments about the probability that two aircraft will "conflict" – get too close horizontally or vertically. In the enroute environment, conflicts are defined as 5 nautical miles (nm) or less of horizontal separation. To make conflict judgments, pilots monitor the progress of their own aircraft (the "ownship") and another aircraft (the "traffic") using an egocentric traffic display. This task is difficult because it requires predicting future aircraft positions given uncertain cues from an IA system about the aircrafts' current positions, speeds, and headings and then making predictions about their distance of separation.

IA designed to assist this task could provide varying levels of support to help the human make this judgment. Support at the "comprehend" level could include descriptive statistics about position, speed, and heading values over a certain time window. The IA could also integrate these cues and make its own assessment about the probability of a conflict, support at the "assess" level. Further, it could provide additional information to the human about the process it used to calculate its judgment, support at the "explain" level.

When considering the cost of development and deployment of IA to support air traffic control tasks, it would be useful to have a detailed framework to characterize this judgment task and enable investigating different design issues. The specific objective of this study was to implement the ELMA framework to investigate the effect of providing additional levels of IA support, in tandem with an automated judgment, on human judgment achievement and consistency. This research was useful as the first empirical use of the ELMA framework (including the taxonomy of judgment support) and to gather empirical data regarding the impact of IA support on judgment performance.

4.2. Methods

Results from a part-task, desktop air traffic control simulator experiment were analyzed. The primary objective was to assess whether the ELMA model could provide useful insight into human-automation judgment performance. We predicted there would be variance in judgment performance due to providing participants with different combinations of IA support levels. By analyzing this variance, we could determine to what degree the ELMA framework could be used to investigate automation design hypotheses.

4.2.1. ELMA conceptualization of the judgment task

Participants were asked to monitor the progress of the ownship and the traffic using a simulated egocentric traffic display. The ownship was flown by an autopilot, so the participants did not need to fly the aircraft. The ownship's speed, altitude, and heading remained constant while uncertainty (sensor noise) was introduced into the speed, lateral position, and heading of the traffic aircraft. The air traffic simulation used for the task was adapted from Bass and Pritchett [7].

To make a probability of conflict judgment one must predict the distance between two aircraft at their point of closest approach, which is a function of the position of the aircraft, the relative heading, and the speed of the aircraft at the time of judgment. Because this was a simulated environment, the environmental criterion and true cues were known. The available cues were simulated to be different than the true cues due to sensor noise. The displayed cues directly resulted from available cues with no added noise as described below.

The IA (Traffic Conflict Prediction System) was instantiated to support the human judge at various combinations of support levels. One instantiation was at the “perceive” only level. The IA transformed the available cues to displayed cues using an egocentric display (see Figure 16 with “perceive” level, “P”, descriptions on the left). The display contained a green aircraft icon representing the position of ownship in the center and a yellow triangle representing position of the traffic. Concentric circles around the ownship represented distances of 5, 10, 20, 30, 40 and 50 nm. A compass was displayed at the 40 nm circle. The heading of ownship was displayed on the compass and its speed was displayed under the green aircraft icon. The traffic triangle pointed in the direction of traffic heading and this was also displayed on the compass with a yellow hash mark. The traffic speed was displayed next to the traffic icon. The ownship and traffic were always at the same altitude. Traffic data were updated once a second. The speed and heading of the ownship remained constant. We can use the ELMA extension to the lens model to represent this judgment task at the “perceive” level of IA judgment support. This is depicted in Figure 17.

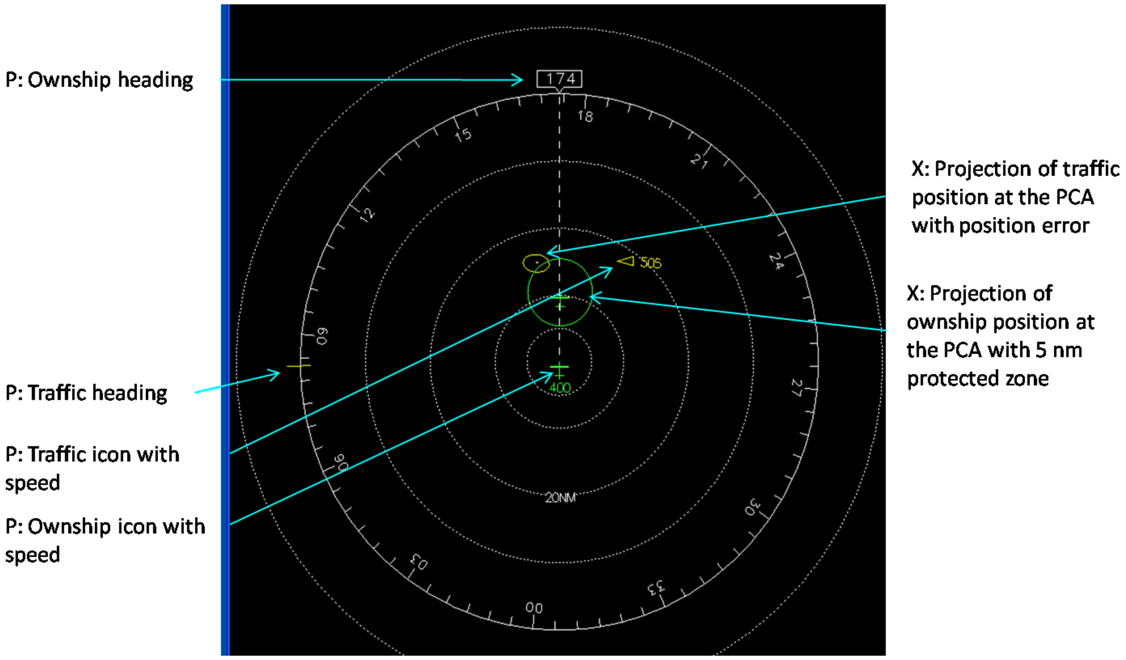

P: Ownship heading
174
15
18
21
12
24
09
27
06
30
03
00
33
505
400
20NM
X: Projection of traffic position at the PCA with position error
X: Projection of ownship position at the PCA with 5 nm protected zone
P: Traffic heading
P: Traffic icon with speed
P: Ownship icon with speed

Figure 16. IA display for air traffic conflict task showing “perceive, P” (left) and “explain, X” (right) support levels.

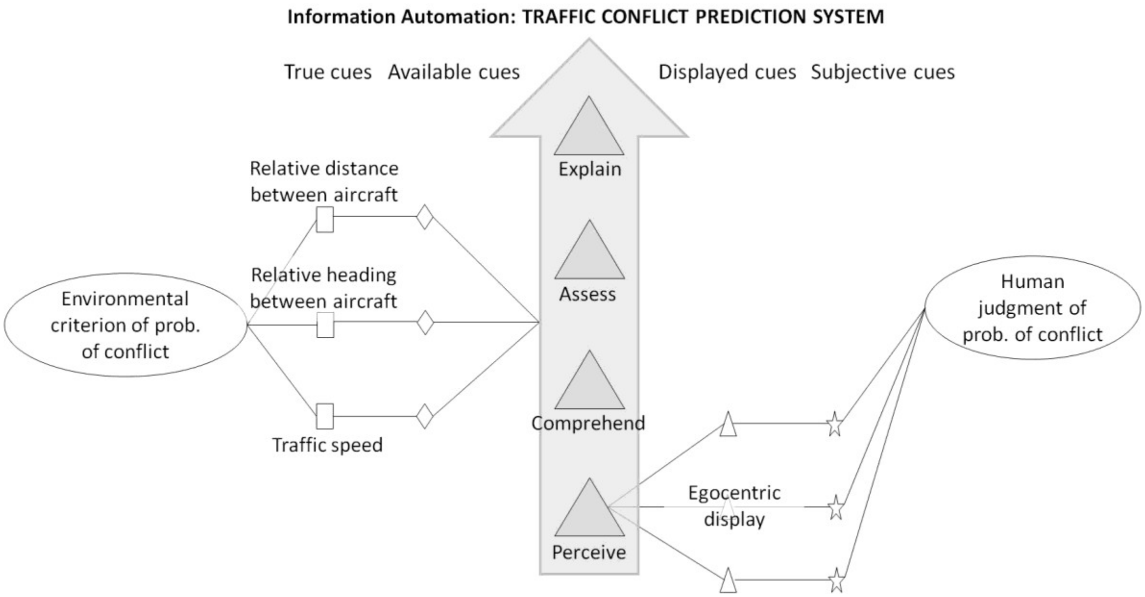

Information Automation: TRAFFIC CONFLICT PREDICTION SYSTEM
True cues
Available cues
Displayed cues
Subjective cues
Explain
Assess
Comprehend
Perceive
Relative distance between aircraft
Relative heading between aircraft
Traffic speed
Environmental criterion of prob. of conflict
Human judgment of prob. of conflict
Egocentric display

Figure 17. ELMA lens model representation of probability of an air traffic conflict judgment task with IA providing “perceive” (P) level of judgment support.

The IA was also instantiated to support the human judge at both the “perceive” (as described above) and “assess” levels of judgment support. To support the human at the “assess” level, the IA calculated probability of conflict judgments by first projecting both ownship and traffic positions to the predicted point of closest approach and then calculating the predicted horizontal miss distance. This was calculated using the available locations, heading, and speed of the ownship and traffic at the time the calculation was made. The probability of conflict was then determined from the cumulative distribution function of the predicted horizontal miss distance with the distance as the mean and its variance calculated as a function of the uncertainty in the lateral position, speed, and heading (see [103] for more details). A slide bar above the display in Figure 16 (not shown in the figure) was used to indicate the IA’s judgment. The ELMA extension to the lens model to represent this judgment task at the “perceive” and “assess” level of IA judgment support is depicted in Figure 19.

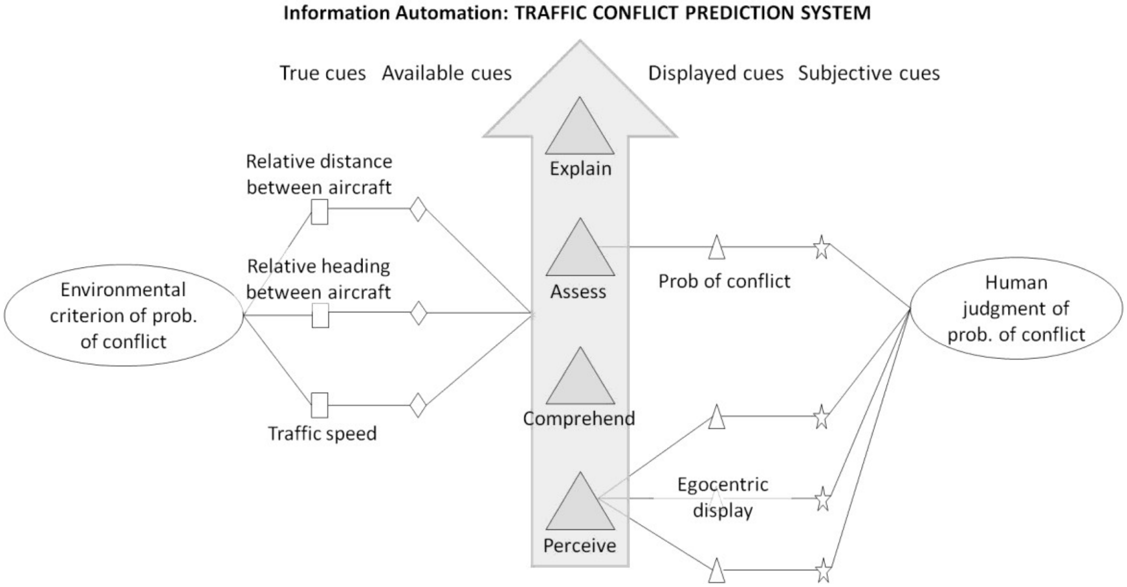
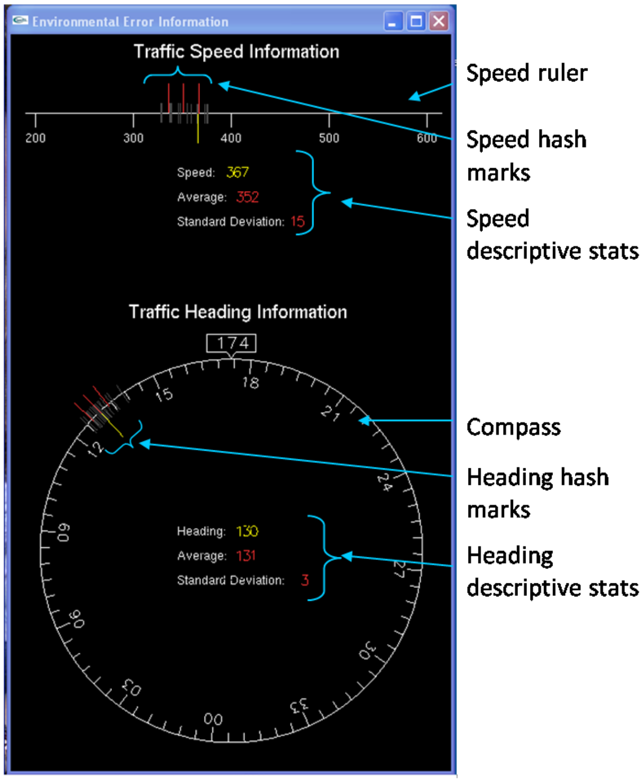

Information Automation: TRAFFIC CONFLICT PREDICTION SYSTEM
True cues
Available cues
Displayed cues
Subjective cues
Relative distance between aircraft
Relative heading between aircraft
Traffic speed
Environmental criterion of prob. of conflict
Explain
Assess
Comprehend
Perceive
Prob of conflict
Egocentric display
Human judgment of prob. of conflict

Figure 18. ELMA lens model representation of probability of an air traffic conflict judgment task with IA providing “perceive” and “assess” (PA) levels of judgment support.

The IA was also instantiated to support the human judge at the "perceive" (as described above) "comprehend" and "assess" (as described above) levels of judgment support. To support the human judge at the "comprehend" level, the IA used an additional display (Figure 19) that contained representations of descriptive statistics of the cues of relative heading and traffic speed. Traffic speed information (top of Figure 19) included a speed ruler which contained a grey hash mark for every traffic speed during a judgment trial (displayed traffic speeds). The speed of the traffic at the time a judgment was required was a yellow hash mark on the speed ruler. The average and standard deviation of the displayed traffic speeds were represented with three red hash marks. The middle hash mark represented the average, and the outer red hash marks represented one standard deviation above and below the average speed. The final speed, average, and standard deviation were also displayed numerically below the speed ruler.

Relative heading information (bottom of Figure 19) contained a compass that had a grey hash mark for every displayed traffic heading during a judgment trial. The ownship heading remained constant and this was indicated at the top of the compass. A yellow hash mark on the compass indicated the final heading of the traffic. The average and standard deviation of the displayed traffic headings were calculated and shown using red hash marks as with the average and standard deviation of the speed (average in the middle, standard deviation marks to the left and right of the average). The heading when the trial ended, along with the average and standard deviation of the displayed traffic headings, were also displayed numerically in the center of the compass. We can use the ELMA extension to the lens model to represent this judgment task at the "perceive," "comprehend," and "assess" levels of IA judgment support. This is depicted in Figure 20.



Speed: 367
Average: 352
Standard Deviation: 15

Heading: 130
Average: 131
Standard Deviation: 3

Figure 19. IA display for air traffic conflict task showing “comprehend” support level.

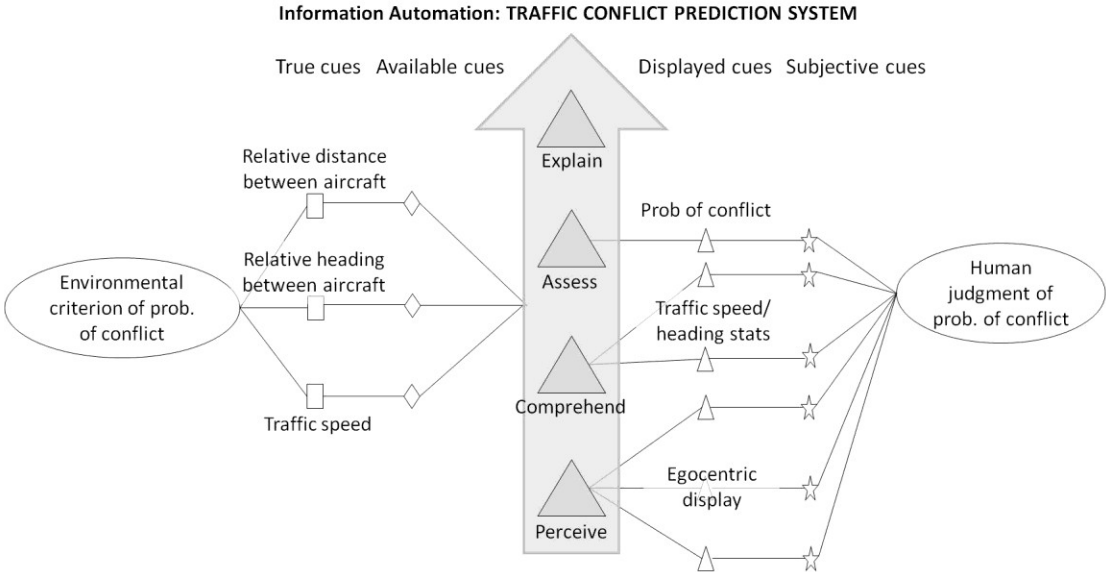

Information Automation: TRAFFIC CONFLICT PREDICTION SYSTEM
True cues
Available cues
Displayed cues
Subjective cues
Explain
Assess
Comprehend
Perceive
Relative distance between aircraft
Relative heading between aircraft
Traffic speed
Environmental criterion of prob. of conflict
Prob of conflict
Traffic speed/ heading stats
Egocentric display
Human judgment of prob. of conflict

Figure 20. ELMA lens model representation of probability of an air traffic conflict judgment task with IA providing “perceive,” “comprehend,” and “assess” (PCA) levels of judgment support.

The IA was also instantiated to support the human judge at all levels of judgment support. To support the human at the “explain” level, the IA also displayed automation strategy information related to how it made its probability judgment: the projected positions of the ownship and traffic. Ownship was represented with a green airplane icon surrounded by a circular green 5 nm protected zone. Projected traffic was represented with a yellow dot surrounded by a yellow two standard deviation position error ellipse representing how the noisy input data (lateral position, speed, and heading) affected the projected location of the traffic at the point of closest approach and thus the automation’s probability of conflict judgment. This level of judgment support is shown in Figure 16, indicated by the annotations to the right of the figure. The ELMA extension to the lens model to represent this judgment task at all levels (including “explain”) of IA judgment support is depicted in Figure 21.

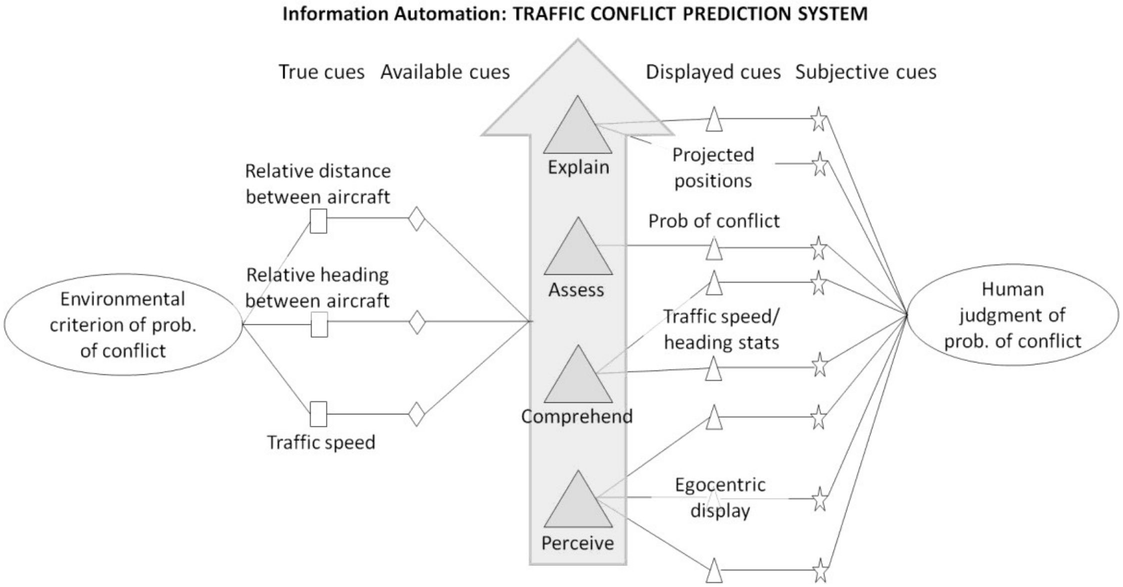

Information Automation: TRAFFIC CONFLICT PREDICTION SYSTEM
True cues
Available cues
Displayed cues
Subjective cues
Explain
Projected positions
Relative distance between aircraft
Prob of conflict
Assess
Relative heading between aircraft
Environmental criterion of prob. of conflict
Traffic speed/ heading stats
Human judgment of prob. of conflict
Comprehend
Traffic speed
Egocentric display
Perceive

Figure 21. ELMA lens model representation of probability of an air traffic conflict judgment task with IA providing “perceive,” “comprehend,” “assess,” and “explain” levels (PCAX) of judgment support.

4.2.2. Judgment profiles

180 judgment profiles were used in the experiment. They were grouped into 6 sessions of 30 profiles each. Each judgment profile consisted of the ownship in the center of the display and one traffic aircraft which flew at one of six possible headings (+/- 45, +/- 90, and +/- 135 degrees from ownship's heading) and five possible speeds (at the same IAS as ownship, +/- 50 knots, +/- 100 knots). The lateral position, airspeed, and heading errors for the traffic aircraft were normally distributed and had standard deviations of 500 meters, 15 knots, and 3 degrees respectively. Every second, new position, speed, and heading errors were added to the actual position, speed, and heading of the traffic aircraft and then displayed, resulting in noisy input data for the human and automation to make judgments.

4.2.3. Participants

Thirty-two male undergraduate engineering students ranging in age from 20 to 23 participated in the experiment. All participants were familiar with the use of computers and had no previous experience with the judgment task.

4.2.4. Procedure used in data collection

Prior to the start of data collection, each participant completed 180 practice profiles (not included in this analysis). Then, each participant completed two days of three sessions each day with thirty profiles in each session. For each profile, participants made two judgments. First, they were provided with IA support at the "perceive" level (Figure 16 and Figure 17) to monitor the ownship and traffic. After a random amount of time (uniformly distributed between 15 and 30 seconds), the screen froze and participants made a probability of traffic conflict judgment at this level of support. Once they made their first judgment, they were then provided with additional judgment support from the IA (with the screen remaining frozen). They then provided a second probability judgment. The trial then continued in faster than real-time so the participant could view the point of closest approach.

4.2.5. Independent variables

All participants were supported at the "perceive" level for their first judgment of every profile (the egocentric display). After making this judgment, each participant was provided with additional IA support in only one of four conditions:

Table 8. Conditions of IA support.

IA level combination	**Description**
A	An automated judgment of probability of conflict
CA	Descriptive statistics of traffic speed and heading over time and an automated judgment of probability of conflict
AX	An automated judgment of the probability of conflict and an explanation of automation strategy information (prediction of ownship and traffic at point of closest approach)
CAX	Descriptive statistics of traffic speed and heading over time, an automated judgment of the probability of conflict, and an explanation of automation strategy information (prediction of ownship and traffic at point of closest approach)

Session (group of thirty trials) was also treated as an independent variable in this study.

4.2.6. Dependent variables

For every judgment profile, the following data were collected: the environmental criterion, true cues, available cues, the human's initial judgment, the human's judgment with additional IA support, and the IA's judgment.

Since all criteria and judgment values were probability estimates, they ranged from 0 to 1. The following transformation was used to stabilize the variance since the tails tended to fall off sharply [114]. This transformation was applied to all criteria and judgments values and used for all subsequent analysis.

$$y = sin^{-1}(\sqrt{x}) \tag{4.1}$$

From these data, derived measures based on the ELMA framework were calculated (see Table 5). The following ELMA measures were derived to characterize the context in which the human made judgments:

- True environmental predictability, $R_{E.t}$
- Available environmental predictability, $R_{E.a}$
- Automation potential, $V_{a.t}$

In this experiment, the displayed environmental predictability, $R_{E.d}$, is equal to the available environmental predictability. Similarly, the true accuracy of the automation, $V_{d.t}$, is equal to the automation potential and the displayed accuracy of the automation, $V_{d.a}$ is subsequently equal to one. Thus, these three ELMA measures are not discussed and we assume that the algorithms used to transform the available cues into displayed cues are an insignificant factor in the performance of the overall human-automation judgment system.

Because in this experiment, every participant was also provided with IA support at the "assess" level, automation's judgment performance (based on available cues) was also characterized using ELMA measures:

- Automation achievement, r_{aA}
- Automation cognitive control, $R_{A.t}$
- Automation linear knowledge, $G_{A.t}$

To characterize the human judge (and thus, the human-automation system) the following ELMA measures were calculated. (Note that participants were not asked for their interpretation of the displayed cues, to collect subjective cues. Thus, we cannot calculate the human's subjective cognitive control, $R_{H.s}$ or the subjective accuracy of the human judge, $V_{d.s}$.)

- Initial human-automation judgment achievement, r_{a1}
- Initial displayed cognitive control, $R_{H.d1}$
- Initial displayed linear knowledge, $G_{H.d1}$
- Additional support human-automation judgment achievement, r_{a2}
- Additional support displayed cognitive control, $R_{H.d2}$
- Additional support displayed linear knowledge, $G_{H.d2}$

4.2.7. Data analysis

Four steps were followed using the ELMA framework. First, the environmental context was characterized (for each session) by modeling the criterion using the true and available cues and then computing the relevant ELMA measures. Second, the IA's performance at the "assess" level was evaluated (for each session) by modeling the automated judgment using the true and available cues and then computing the relevant ELMA measures. Third, linear models were created for each participant, for each session, using the displayed cues and then computing the relevant ELMA measures. Last, nomothetic analysis of the ELMA measures for IA support and session effects was conducted.

The experimental design for nomothetic analysis was a repeated measures, mixed model design. IA support condition, session, and the session-IA support interaction were fixed effects. Participants were nested within IA support condition and were treated as a random effect in the model. Post hoc analysis was conducted using Tukey's Honestly Significant Difference (HSD). Wilcoxon Signed Rank tests were also used to determine if differences existed between initial human judgment achievement and achievement with increased IA support.

All derived measures from the ELMA framework are correlations. Therefore, before performing the nomothetic data analysis described above, the correlations were transformed using Fisher's r to z_r transformation (equation 4.2) to obtain normally distributed variables as suggested by Cooksey [40]. However, the measures are reported and graphed prior to transformation.

$$z_r = \frac{1}{2} log_e \left(\frac{1+r}{1-r}\right) \tag{4.2}$$

4.3. Results

The results of this experiment are presented using α = 0.05 for significance and α = 0.1 for a trend. The environmental context and automation performance are described first. This is followed by an investigation of effects of IA support and session on human-automation judgment performance. The session-IA support interaction was never significant and is therefore not reported.

4.3.1. Environmental context and automation performance

To characterize the context in which the human must make judgments, we can start by deriving the true and available environmental predictability for each session of the experiment. These measures indicate how predictable the criterion is based on the cues, which essentially represents an upper bound on the judgment performance of the human judge.

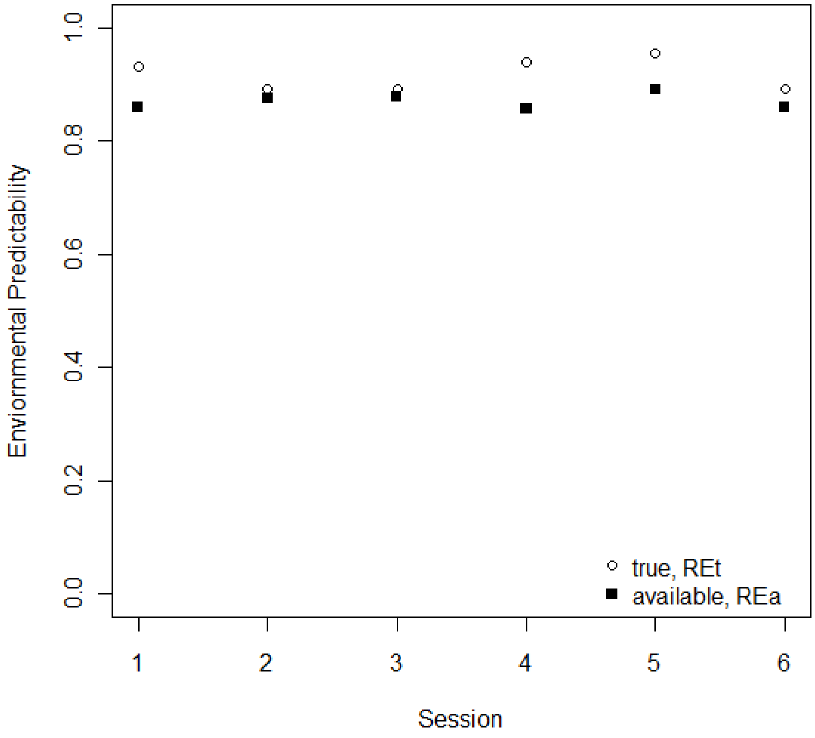


Figure 22. True and available environmental predictability over sessions.

As defined by ELMA, the ratio of true and available environmental predictability is defined as the automation potential. For this experiment, the automation potential ranged from 0.91 to 0.98, characterizing the transformation from true to available cues. The true cues are not perfect predictors of the criterion (as we would theoretically expect). This is because the judgment task is a prediction of a future environmental state and the true cues used to model the criterion are from the time when the judgment is made.

For each session, we can also characterize the performance of the IA in its ability to provide judgment support at the “assess” level (i.e., provide its own judgment about the probability of conflict). In all conditions, the participants were provided with more support than just the “assess” level.

However, when characterizing the judgment context that the human had, it is also useful to understand how the automation is performing and contributing to the context. For example, if the automation was not a good judge of the criterion and it was presenting the human with this information, this may negatively influence the human’s performance. From Figure 23 we can see that the automation was not a perfect judge of the criterion. Its judgment achievement ranged from 0.92 to 0.97, with an average judgment achievement of 0.94. However, because we know that the environmental criterion is not perfectly predictable with the true cues, the automation’s judgment performance may still be acceptable.

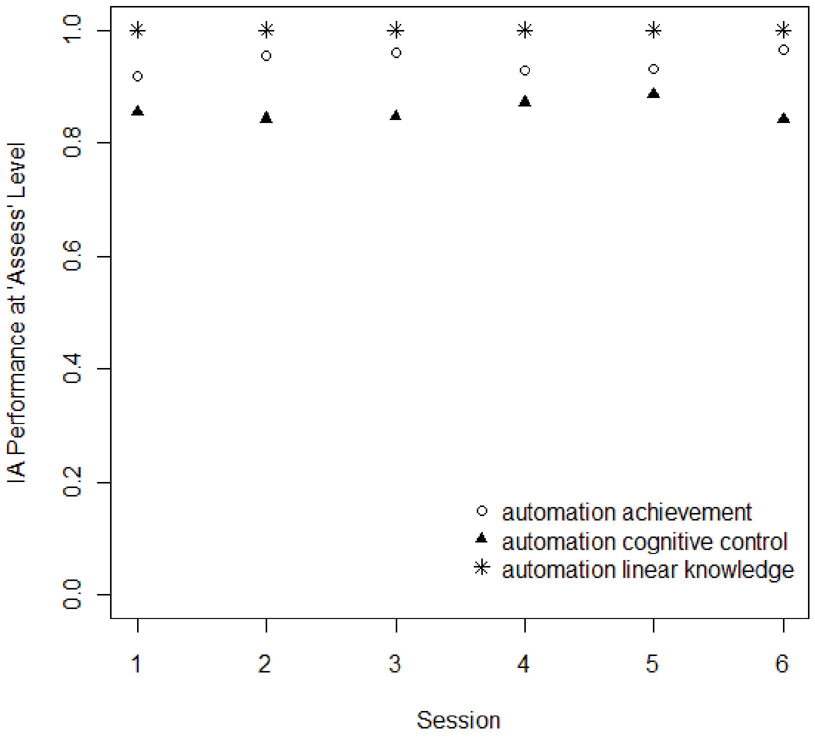


Figure 23. Automation performance at the “assess” level of judgment support.

4.3.2. Nomothetic results of judgment performance

Six linear models were created for each participant (one for each session) and ELMA measures were subsequently calculated using each model. Figure 24 and Figure 25 contain the initial average human-automation judgment achievement (r_{a1}) grouped by session and IA support condition respectively. There was no significant difference between the participants assigned to the four IA support conditions in their initial judgment achievement (as we would expect). However, session was significant ($F_{5, 168} = 4.8635$, $p < 0.001$) (see Figure 24).

Post hoc analysis using Tukey's HSD indicates the only significant difference between sessions for initial judgment achievement was between sessions 3 and 4 ($p < 0.001$) and between sessions 3 and 6 ($p < 0.001$). The decrease from 3 to 4 may be explained by the fact that after session 3, the participants had a break before starting session 4 the next day. Participants may have needed a session to become re-familiarized with the task and the output from the automation.

Overall, the participants' second judgment achievement with additional support from the IA was better than their initial judgments at the "perceive" level only. A Wilcoxon Signed Rank test indicates that average second judgment achievement, r_{a2} ($\mu = 0.91$, $\sigma = 0.06$) was significantly higher than the average initial judgment achievement ($V = 0$, $p < 0.001$) and much closer to the automation's judgment achievement ($\mu = 0.94$, $\sigma = 0.02$). Additionally, using Levene's test, the variance of second judgments across all conditions was significantly smaller compared to the variance of initial judgments ($F_{191, 191} = 10.251$, $p < 0.001$) and the floor of judgment achievement was raised from -0.16 (initial) to 0.66 (increased IA support) across all participants. Thus, we can conclude that adding any combination of support levels above "perceive" enhanced human-automation judgment performance in terms of judgment achievement (or correlation with the environmental criterion).

Further, IA support condition ($F_{3, 168} = 5.979$, $p = 0.001$) and session ($F_{5, 168} = 9.5782$, $p < 0.001$) both had a significant impact on participants' second judgment achievement. Figure 24 and Figure 25 also contain the second judgment achievement (r_{a2}) by session and IA support condition. Tukey's HSD post hoc analysis for the session effect on judgment with additional IA support indicates that sessions 2, 3, 5, and 6 were significantly higher than session 1 and also that session 4 had a trend to be higher than session 1. This could indicate that participants needed a session (session 1) before they fully understood how to use the support from the IA to improve their judgments.

More interestingly, Tukey's HSD analysis also indicates that the added A level had significantly lower achievement than both the added CA (p = 0.003) and CAX (p = 0.02) conditions (see Figure 25). There was also a trend for the AX condition to be lower than the CAX condition (p = 0.08). There was no significant difference between the CA and CAX conditions implying that for this particular study, judgment support at the "explain" level did not significantly improve judgment achievement compared to when participants were already provided with support at the "comprehend" level. These results may indicate that the "comprehend" level is primarily responsible for the enhanced human-automation judgment achievement.

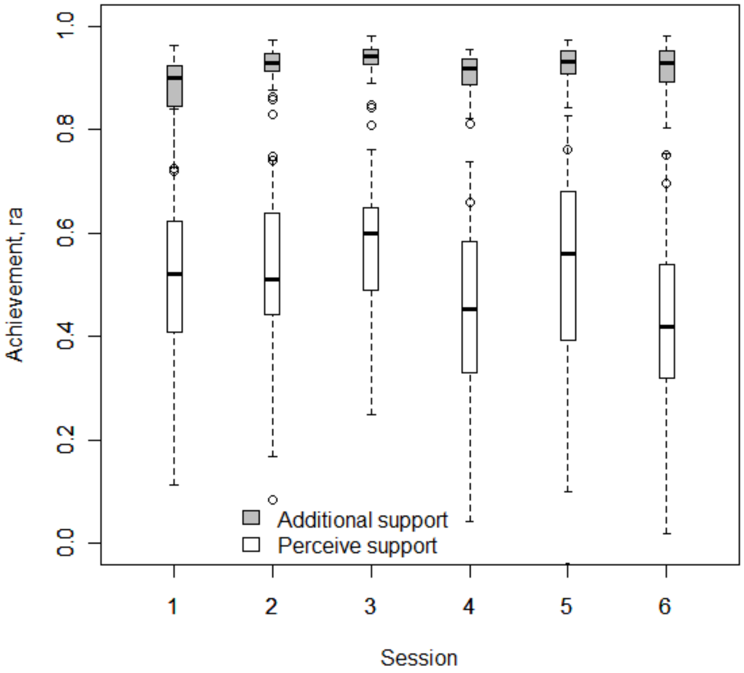


Figure 24. Initial achievement with “perceive” support and achievement with additional IA support by session (n = 32 for each session).

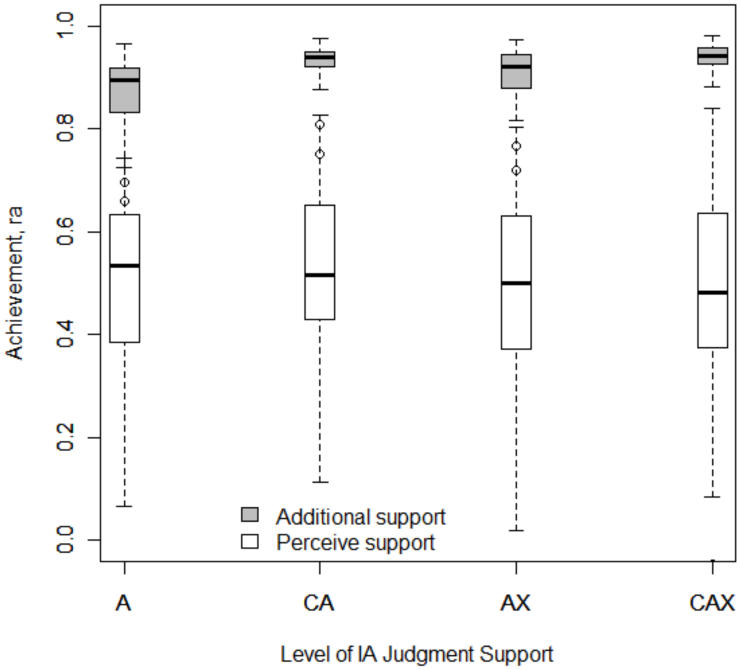


Figure 25. Initial achievement with “perceive” support and achievement with additional IA support by IA support condition (n = 8 for each support condition).

Figure 26 and Figure 27 contain initial and second displayed cognitive control (R_{Hd}) grouped by session and IA support condition respectively. Overall, the participants’ cognitive control of their second judgments with additional support from the IA was better than their initial judgments at the “perceive” level. A Wilcoxon Signed Rank test indicates that average second judgment cognitive control, $r_{H.d2}$ (μ = 0.86, σ = 0.05) was significantly higher than the average initial cognitive control, $r_{H.d1}$, (μ = 0.46, σ = 0.18) (V = 0, p < 0.001).

There was no significant difference between the participants assigned to the four IA support conditions in their initial cognitive control (as we might expect). However, their cognitive control with increased IA support was trending to be impacted by IA support condition ($F_{3, 168}$ = 3.25, p < 0.1). Session was also significant for both initial judgments ($F_{5, 168} = 3.69$, $p < 0.01$) and judgments with increased IA support ($F_{5, 168} = 35.82$, $p < 0.0001$).

Post hoc analysis using Tukey's HSD indicates a trend for the CA and CAX conditions to be higher than the A condition ($p = 0.08$ and $p = 0.09$ respectively) (see Figure 27). This may imply that the added support at the comprehend level may have aided participants in more consistently applying their judgment strategies across judgment profiles.

Post hoc analysis using Tukey's HSD also indicates the only significant differences between sessions for initial cognitive control was that session 6 was significantly lower than sessions 5 ($p < 0.002$), 3 ($p < 0.02$), and 2 ($p < 0.04$). This decrease in cognitive control could have been due to participant fatigue with the experiment by the last session. Post hoc analysis on the second judgment cognitive control measure indicates that sessions 6 and 1 were significantly lower than the other sessions and session 5 was significantly greater than sessions 2, 3, and 4. Again, the decrease in the last session may have been due to participant fatigue. The low cognitive control in the first session may be the result of participants still getting used to the increased automation support and they had not yet adapted their judgment policies to reflect this added aid.

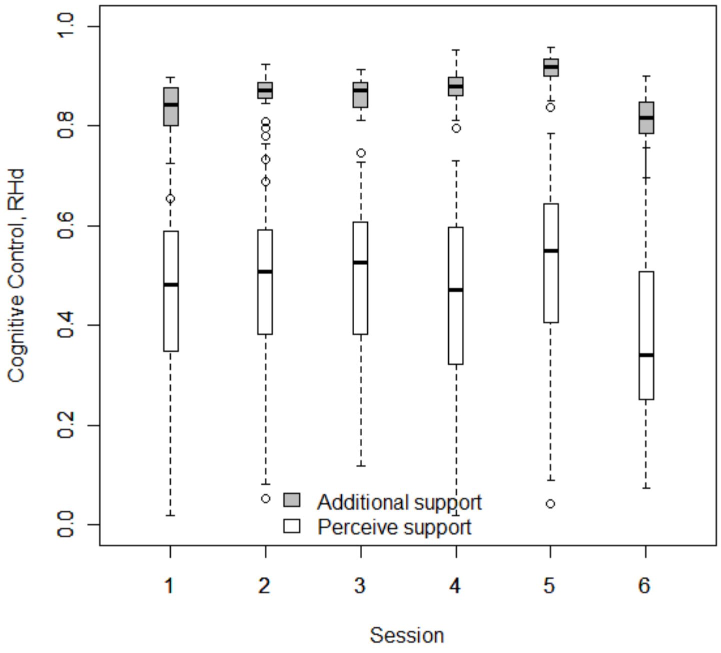


Figure 26. Initial displayed cognitive control with “perceive” support and displayed cognitive control with additional IA support by session.

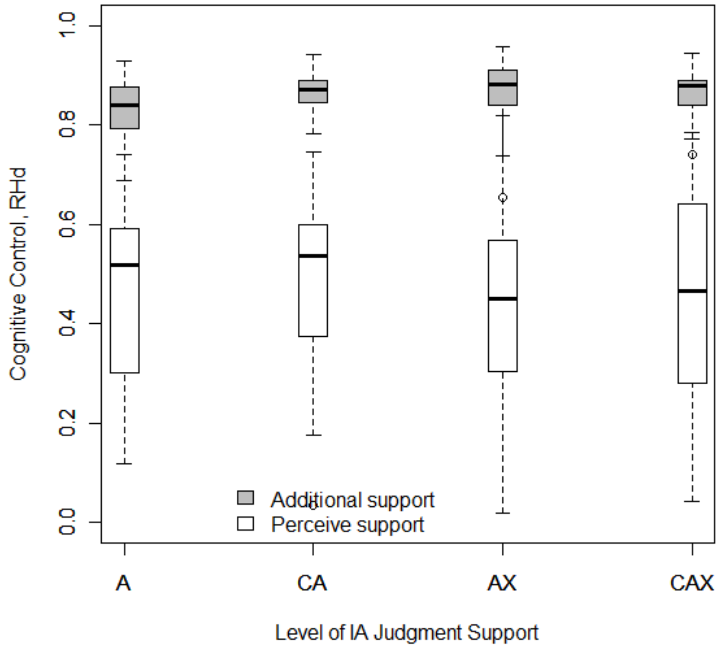


Figure 27. Initial displayed cognitive control with “perceive” support and displayed cognitive control with additional IA support by IA support condition.

There were no significant effects on displayed linear knowledge due to either session or IA support condition for either initial judgments at the P level or for second judgments with increased IA support. There was also no difference in un-modeled agreement between participants’ initial judgments and their judgments with additional IA support.

Recall from the ELMA lens model equation (equation 3.32) that human-automation judgment achievement is dependent on displayed linear knowledge ($G_{H.d}$), displayed environmental predictability ($R_{E.d}$), and displayed cognitive control ($R_{H.d}$). From our analyses, we know that the low judgment achievement (μ = 0.49, σ = 0.19) for participants’ initial judgments was primarily impacted by low displayed cognitive control (μ = 0.46, σ = 0.18). However, their second judgment achievement (μ = 0.91, σ = 0.06) was driven by both displayed cognitive control (μ = 0.86, σ = 0.05) and available (and displayed) environmental predictability ($\mu = 0.87$, $\sigma = 0.01$). This implies that the additional IA support improved displayed cognitive control to a level that was similar to the predictability of the environmental criterion.

4.4. Discussion

This was the first instantiation of the ELMA framework. We used ELMA to represent and describe the judgment task and the levels of IA support under which design hypotheses were derived and tested. ELMA also guided the analysis of judgment context and performance. This analysis demonstrated that the ELMA framework, using multiple linear regression models, were sufficient to investigate the effects of varying conditions of IA support on judgment performance.

4.4.1. Effect of automated support on human-automation judgment performance

This research sought to investigate the impact of IA support on human judgment performance. When provided with IA support above the “perceive” level, judgment achievement significantly improved for all participants across all sessions and all added IA support conditions. This shows the value of added automation support on the human’s judgment for this task. Additionally, the variance of the participants’ second judgment was reduced compared to the initial judgment indicating that the increased support reduced variability and allowed for a range of operators with varying capabilities to maintain similar performance ranges. The increased IA support also raised the overall floor of judgment achievement to support absolute levels of performance (see Figure 25) and increased participants’ overall cognitive control (see Figure 27).

Participants provided with additional support at the “comprehend” level (CA or CAX) had significantly higher joint judgment achievement compared to those provided with only the added “assess” level (A). There was no significant difference between judgment achievement by participants in the CA and CAX support groups. This implies that adding automation strategy information did not significantly help when participants were also provided with aid in comprehending the cues for this judgment task. A similar pattern of performance enhancement with the CA and CAX groups was also seen in the participants' cognitive control. The added support at the comprehend level may have contributed to enhanced cognitive control in that participants were able to better understand the environmental context (i.e., the cues) and were thus able to more consistently apply their judgment strategies, resulting in greater judgment achievement. Similar results were found by Bisantz et al. [19] when they found that individual differences in judgment performance were most attributed to differences in cognitive control.

In some cases, total system performance (as measured by second judgment achievement with increased IA support) was greater than either the human's initial judgment achievement or the automation's judgment achievement alone. One participant in the CA condition and one participant in the AX condition had greater judgment achievement than the automation's judgment achievement for five of the six sessions. These participants developed a sophisticated strategy in which they were incorporating the IA support at both the "comprehend" or the "explain" level into their judgments, rather than simply adapting to the automation's output at the "assess" level. This is similar to behavior described in Bass and Pritchett [7] where a participant was able to use the automation's output in his judgment when it was of more value and to ignore the automation otherwise.

4.4.2. Implications for the design of automation

These results have implications for automation design. In this experiment, the displayed cues' uncertainty was tied directly to the uncertainty between the true cues and the available cues in the environment (imperfect sensors providing information regarding the traffic speed and heading). Although providing participants with support regarding the automation's judgment strategy may improve performance compared to those receiving only an automated judgment (A vs. AX), this level of support does not appear as beneficial to participants as automation support pertaining to comprehending the cues (A vs. CA).

It is possible that automation support comprehending the cues allowed the participants to exploit the automation more effectively as they understood how the automation's judgment achievement varied based on factors in the environment. This result supports the human judgment literature related to the impact of cognitive feedback on judgment performance. In a review of over 20 human judgment experiments involving cognitive feedback, Balzer et al., [105] found that environment information is the component of feedback with the greatest effect on human judgment performance. Environment information (graphical and statistical information related to the predictability of the environmental criterion, cue weights, and the function form relating the cues to the criterion) also increased human judgment performance in a baseball prediction task over other forms of cognitive feedback [106]. Further, both trained and untrained judges also performed better with environment information during a medical diagnosis task [115]. Thus, for some judgment tasks, it may be that humans may not necessarily need to understand the underlying algorithm(s) used by the automation if they understand how the automation performs under different conditions, similar to results found in [72]. It is likely that when designing information analysis automation to be used in noisy environments where the automation's judgment achievement is correlated with noisy input data, it is better to show additional environment information than the automation's judgment strategy. However, research should investigate where this pattern no longer holds (i.e. where environmental noise increases, reducing the automation's judgment achievement as well as placing a ceiling on human judgment performance given the provided cues, regardless of the environmental information provided).

In this experiment, the automation's judgment achievement at the "assess" level was high (which was provided to all participants) compared to the participant's initial judgment achievement when only provided with support at the "perceive" level. However, if this were not the case, the amount of IA support could have had a different effect on judgment performance. For example, in a different study, when participants could observe explanations of why automation was making errors, they tended to increase adaptation to the automation, even when unwarranted [116] and particularly when trust of the automation exceeded self-confidence [117]. High automation error rates have also resulted in lower subjective measures of trust in automation in numerous studies [118], particularly when participants are aware of conditions affecting the automation's reliability [72]. Trust, in turn, could impact how the human interacts with the automation at different support levels.

There are numerous ways to represent the displayed cues and the support of the automation. Thus, in order to fully generalize the results found here, further research using different representations in different domains should be conducted. Additionally, it would be interesting to investigate the benefit of providing added support at the "comprehend" level, without including an automated judgment (the "assess" level) for the participants to consider. Another limitation of this study is that the participants were undergraduate students performing a simplified air traffic conflict prediction task with no secondary tasks to perform. Automation support conditions should also be tested with trained operators in more naturalistic environments. In particular, it is unclear if the benefit of additional support from the automation would hold in settings where conflict detection is only one of many tasks to perform. Also, it may be that experienced participants may have been better able to understand the strategy information, or alternatively, have produced better independent judgments and therefore relied less on the automation.

5. Applying ELMA to a healthcare quality judgment task

The chapter discusses how the ELMA framework was used to understand the health care judgment task of assessing population-based quality of care and to inform the design of an automation system to support such judgment tasks. This application of ELMA also includes an empirical investigation of the impact of automated judgment support levels on performance for quality judgments.

5.1. Introduction

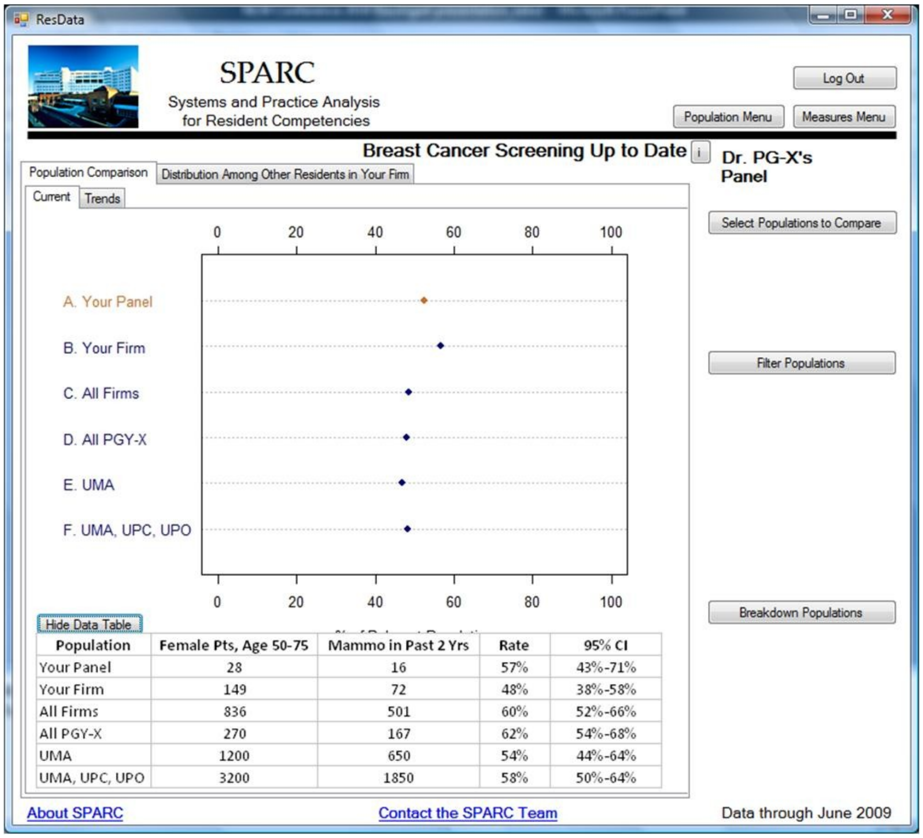
Judgments made to assess the quality of health care based on actual practice data have gained increased attention and have been the focus of many health care initiatives and medical education curriculum efforts. For example, the 2006 Tax Relief and Health Care Act (TRHCA) required the establishment of a physician quality reporting system, including an incentive payment for eligible professionals who satisfactorily report data on certain quality measures [119]. Also, the Accreditation Council for Graduate Medical Education (ACGME) recently specified practice-based learning and improvement (PBLI) as one of the core learning requirements for residency programs in the United States [15]. Residents (physicians who have completed medical school and are undergoing required training in their medical specialty) must demonstrate the ability to investigate and evaluate the quality of their practice behavior in order to identify strengths, weaknesses, and areas for improvement. Further, the American Board of Medical Specialties states that physicians must investigate and evaluate their patient care practices as one of six key elements in its Maintenance of Certification program [120] and the American Board of Internal Medicine mandates that physicians must assess clinical performance with practice reviews to determine compliance with accepted standards and guidelines in order to complete recertification [121].

Despite its importance, physicians are not trained to do this task and have actually shown a limited ability to accurately judge their quality of care based on actual practice data [16]. A number of approaches have been implemented to address the need for supporting quality assessment in health care. These include web-based applications [122], participation in quality improvement projects [123], [124], video-taped review of encounters with standardized patients (actors) [125], and peer chart audits [126]. However, most efforts have focused on walking physicians through the steps of quality improvement initiatives and few have focused on building skills to investigate and evaluate practice data as drivers for quality improvement. Further, in most of the implementations mentioned above, clinical outcome measures were used to indirectly evaluate practice investigation efforts (e.g., how clinical outcomes improved after the quality improvement projects were completed [124]). Although improving clinical outcomes is the overarching goal of quality assessment, it is not clear if the current efforts are improving care based on direct investigation and judgments of practice behaviors.

One strategy for addressing the need to support judgments of quality of care is to use information automation that presents population-based information (i.e., aggregated data, not individual patient data). Such population-based information can facilitate physicians' judgments through analysis of their practice and also by enabling comparisons between other populations, such as those of their peers or to an entire clinic. Investigating populations of patients may also aid in evaluating resident physicians, understanding where to direct quality improvement initiatives or limited resources, and monitoring adherence to guideline recommendations.

Studying and supporting quality of care judgments with population-based information is difficult. First, it is difficult to define an environmental criterion. There is not a consistent and precise definition of how to make judgments regarding quality of care [127] and the judgments vary depending on the specific type of care you are assessing (e.g., preventative medicine, hypertension control, etc.). Second, there are many sources of uncertainty in available cues used to make quality of care judgments to evaluate practice behaviors. For example, patient data may be missing (e.g., due to receiving care at other facilities), erroneous (e.g., due to inaccurate medical records), unavailable (e.g., the pharmacy database does not communicate with a clinical data repository), or confounded by variables outside of the physicians' control (e.g., they recommended a mammography, but the patient failed to go to her appointment because she did not have insurance). Third, few tools currently exist to support quality of care assessments at a population-level and there are limited guidelines for the design of such tools, including what data (or cues) to include and what level of judgment support to provide.

At the University of Virginia (UVa), an IA judgment support tool is currently under development to present population-based patient data to health care providers. The Systems and Practice Analysis for Resident Competencies (SPARC) tool presents population-based reports of various demographic data, outcomes, and process measures to enable resident physicians to investigate their practice behaviors in an exploratory way [17], [23]. For example, residents can view a report of different breast cancer screening up-to-date rates as in Figure 28. Rates for six different populations associated with the resident physician are shown (from top to bottom): "Your Panel" (the population of patients assigned directly to the resident), "Your Firm" (the population of patients assigned to a small group of peer residents that is overseen by one attending physician), "All Firms" (the population of patients assigned to all resident firms at the clinic), "All PGY-X" (the population of patients assigned to all residents at year X in their training), "UMA" (the population of patients assigned to the entire clinic, including those seen by residents and attending physicians), "UMA, UPC, UPO" (the population of patients assigned to three different clinics combined).



Population	Female Pts, Age 50-75	Mammo in Past 2 Yrs	Rate	95% CI
Your Panel	28	16	57%	43%-71%
Your Firm	149	72	48%	38%-58%
All Firms	836	501	60%	52%-66%
All PGY-X	270	167	62%	54%-68%
UMA	1200	650	54%	44%-64%
UMA, UPC, UPO	3200	1850	58%	50%-64%

Figure 28. SPARC report of breast cancer screening up-to-date rates for six populations of patients.

The current version of SPARC allows for exploratory investigation into the quality of different aspects of care. We previously evaluated residents' perceptions of the utility and usefulness of SPARC, including the usability of the graphs and the filtering functionality [128]. Of 30 first year residents and 63 second and third year residents, 21 and 42 completed a questionnaire (70% and 67%, respectively). 98% of respondents agreed or strongly agreed that the graphical comparisons were easy to interpret and 94% agreed or strongly agreed that the graphical comparison helped them understand differences between their patients and others.

However, there are still concerns about the acceptance of SPARC due the lack of structure to support more efficient means of making quality judgments (i.e., moving away from an exploratory tool and providing more directed automated support). Thus, the objective of this research is to apply the

ELMA framework to understand and support quality judgment tasks directed at specific areas of interest (e.g., common diseases). For this application, we have chosen to focus on one major health care quality area: *assessing quality of hypertension care* provided by physicians. This specific quality judgment was chosen because in 2010, high blood pressure cost the United States $76.6 billion in health care services, medications, and missed days of work [129] and more than 50 million people in the United States were diagnosed with hypertension [130]. Although, hypertension is a major health care concern, it is unclear how to assess the quality of hypertension care provided by physicians. Further, few tools allow physicians to investigate different aspects of population-level hypertension care simultaneously, such as blood pressure control and adherence to medication guidelines. We will demonstrate that the ELMA framework and methodology can be applied to understand this novel health care judgment task and to inform the design of future versions of SPARC to support direct quality judgment tasks.

5.2. Methods

Our primary goal of this work was to apply the ELMA framework and demonstrate its usefulness in understanding health care quality judgments supported by automation. Following the ELMA methodology, there were three main objectives. The first objective was to identify the appropriate cues necessary to make judgments of population-level hypertension care. This involved a two-phased approach consisting of a document analysis followed by a focus group.

The second objective was to understand how these cues differentially influence judgment and the third objective was to evaluate the effect of level support on judgment achievement, consistency within individual physicians, and reliability across physicians. These objectives involved developing an apparatus to test design hypotheses. This consisted of creating both judgment profiles and a prototype information automation system to support the judgment task. Three versions of the prototype were instantiated under three conditions of judgment support based on the ELMA framework. A human-subject experiment was then conducted with twenty-four internal medicine resident physicians. Our logic for the third objective was that by analyzing variance in judgment performance induced by the amount of IA support, we could demonstrate the degree to which ELMA is useful to test IA design hypotheses.

Further, we conducted post-hoc analysis of variance in judgment performance between best and worst performing individuals. Our logic for this analysis was that by analyzing human judge-specific variance, we could demonstrate the degree to which ELMA is useful to investigate sources of individual differences in judgment performance. This could have implications for the design of training interventions targeting specific aspects of the human-automation judgment process.

5.2.1. Identifying available cues needed for judgment task

We employed a two-phased method to identify the available cues needed to judge the quality of hypertension care: first a document analysis and then a focus group with internal medicine attending physicians. Because this particular judgment task does not currently take place in most health care systems (including UVa), an objective analysis of the ecology and a verbal protocol analysis were not possible.

We consulted the following clinical guidelines to identify clinical measures (or cues) used to assess hypertension care at a population-level (i.e., beyond the quality of care received by a single patient).

Table 9. Documents analyzed to identify cues for quality of hypertension care judgment.

Name of clinical guideline documents	**Organization that authored guidelines**
Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure	The National Heart, Lung, and Blood Institute
National Voluntary Consensus Standards For Clinically Enriched Administrative Data	National Quality Forum
Physician Quality Reporting Initiative	Centers for Medicare and Medicaid Services
National Quality Measures Clearing House	Agency for Healthcare Research and Quality
The Healthcare Effectiveness Data and Information Set (HEDIS)	National Committee for Quality Assurance

Based on the document analysis, we narrowed down the hypertension quality cues to six, based on availability (i.e., if these measures were captured electronically at UVa). We also only considered those measures applicable to patients between the ages of 18 and 74 and who had a recorded diagnosis of hypertension:

1. Percent of patients without diabetes with most recent blood pressure (BP) at or below 140/90 mmHg
2. Percent of patients with diabetes with most recent BP at or below at 135/85 mmHg
3. Percent of patients with diabetes and a positive microalbumum test on either angiotensin converting enzyme (ACE) or angiotensin receptor blockers (ARB) medications

4. Percent of patients on more than two anti-hypertensive medications, with at least one being a diuretic
5. Percent of patients on ACE, ARB, or diuretic with potassium checked within last 15 months
6. Percent of patients with creatinine checked within last 15 months

Seven internal medicine attending physicians at UVa subsequently participated in a focus group to discuss the cues. Each cue was presented one at a time, including the documented source of the cue. Physicians were given the opportunity to discuss the validity of the cue (based on evidence in the literature) as indicators of quality of hypertension care. The focus group unanimously decided to delete cues 4 and 5 due to opinions of there not being clear evidence in the literature that the cues adequately represented the quality of hypertension care. Cues 1 and 2 were also combined to be: Percent of patients at or below "goal" BP (where goal BP is defined differently for patients with and without diabetes based on 1 and 2 above). Thus, the three cues (one outcome measure and two process measures) needed to judge the quality of hypertension care were determined to be:

1. Percent of patients at or below goal BP (where goal is 140/90 mmHg for patients with a diagnosis of diabetes and is 135/85 mmHg for all other patients)
2. Percent of patients with diabetes and a positive microalbumum test on either ACE or ARB medications (the recommended medications)
3. Percent of patients with creatinine checked within last 15 months

5.2.2. ELMA conceptualization of the judgment task

For a physician to make a quality of hypertension care judgment regarding a specific physician's (e.g., a resident physician's) panel of patients, one must consider the three cues (percent of patients at goal BP, percent of diabetes patients on recommended medications, and percent of patients with creatinine checked) related to the panel and then combine the cues to assess the quality of care. The environmental criterion related to this judgment would be the actual quality of care that that physician is providing. This criterion would be related to the three true cues. However, these true cues would be very difficult to access. For example, to know the true percentage of patients at goal BP, one would need to simultaneously measure each patients' exact BP right at the moment the judgment of quality of care was to take place. Available cues related to the true cues are those that can be acquired from an electronic medical record (EMR) database. For example, a query can be run on the EMR database at UVa to obtain the percentage of patients at goal BP based on the recording of their last BP measurement taken at UVa.

IA can be designed to transform the available cues to displayed cues in any combination of the support levels. For example, IA could be developed at the "perceive" level where the three available cues are obtained from the EMR database and then transformed to displayed cues (possibly as dot plots, similar to Figure 28) for the physician to interpret and make judgment regarding the quality of care. The ELMA representation for this task for this level of IA support is shown below.

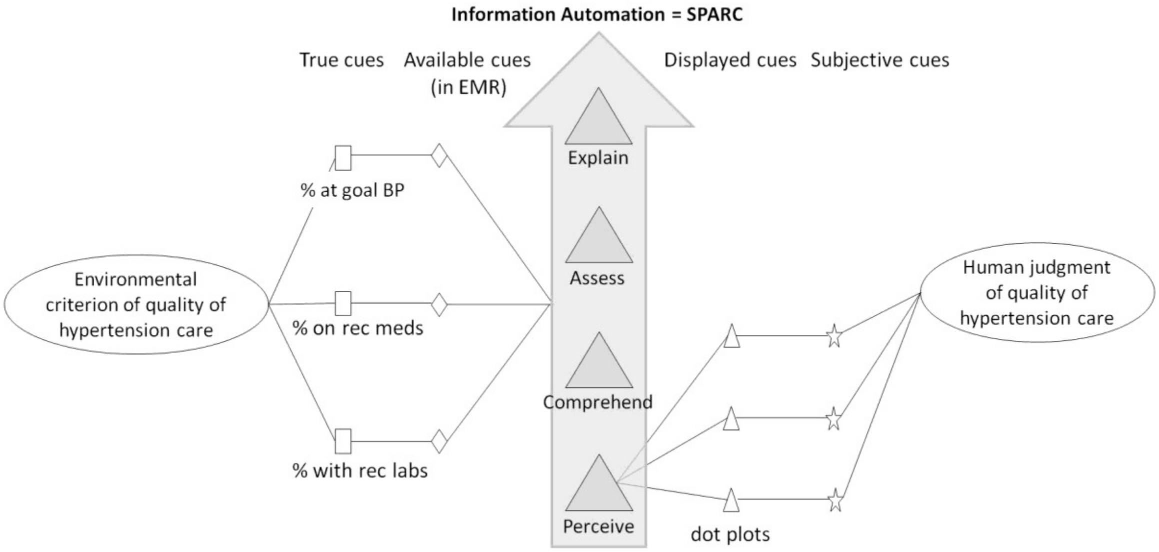


Figure 29. ELMA lens model representation of hypertension quality of care judgment task with IA providing “perceive” (P) level of judgment support.

IA could also be developed to support the physician at the “comprehend” level of support. This could include presenting descriptive statistics of the available cues, such as displaying the percentile rank for each cue compared to a set of cues (i.e., the percentile rank for one physician’s panel of patients compared to all other physicians in the clinic). The ELMA representation for this is shown below.

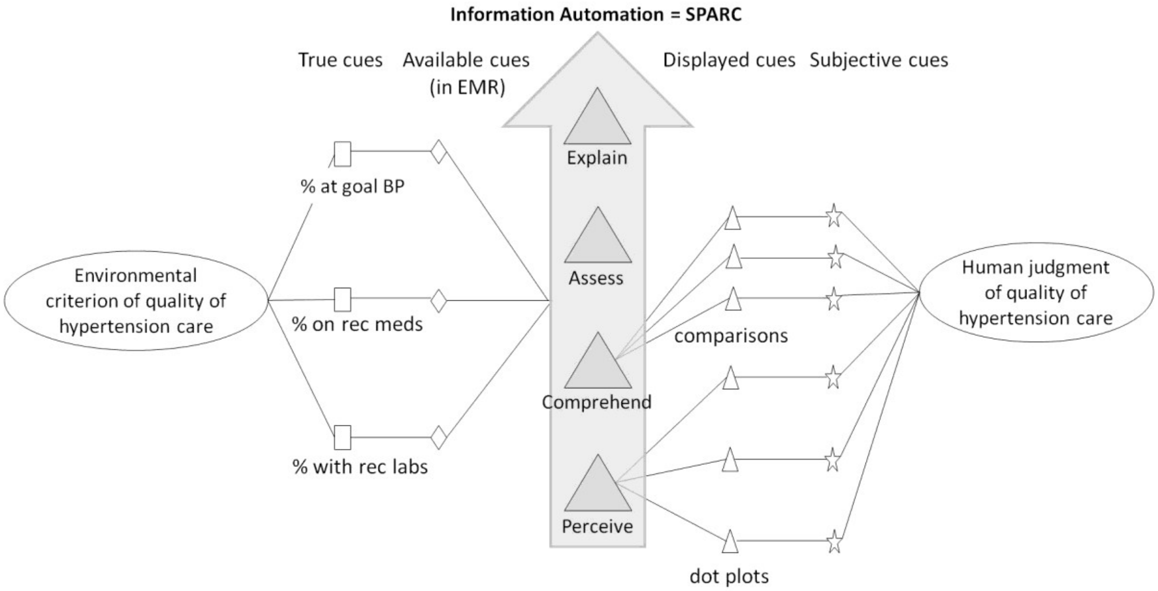


Figure 30. ELMA lens model representation of hypertension quality of care judgment task with IA providing “perceive” (P) and “comprehend” (C) levels of judgment support.

IA could also be developed at the “assess” level where the automation uses an algorithm to combine the three cues and present an automated quality of care judgment to the human judge. For example, the IA could judge the quality of hypertension care by first computing a weighted average on the three available cues and then comparing that average to other physicians’ averages and assign a judgment regarding the quality of care. The ELMA representation for the “assess” level is shown in Figure 31.The IA could also provide support at the “explain” level by indicating its strategy for assessment (as described above).

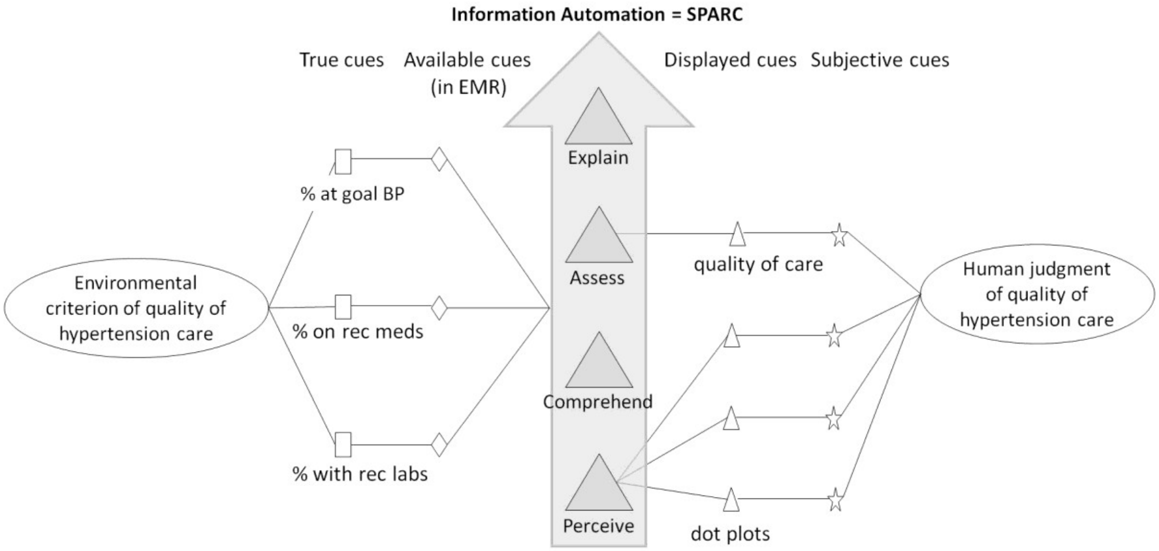


Figure 31. ELMA lens model representation of hypertension quality of care judgment task with IA providing “perceive” (P) and “assess” (A) levels of judgment support.

5.2.3. Hypotheses regarding level of automated judgment support

Results of the first application of ELMA and empirical analysis of the impact of IA judgment support (Chapter 4 of this dissertation) found that human-automation judgment performance was enhanced when IA support included the “comprehend” (C) level (i.e., either PCA or PCAX compared to PA or PAX). However, the benefit of C support was not evaluated without also including support at the “assess” (A) level (i.e., we did not test C, PC, CX, or PCX). For this task it would not make sense to exclude support at a minimum of the P level. The previous study also found no added benefits to providing IA support at the “explain” (X) level. Thus, we have two hypotheses regarding the level of IA support for this judgment task based on these prior results:

1) Support at PC levels will yield the highest level of judgment achievement compared to P only or PA.
2) Support at PC levels will yield the highest level of cognitive control compared to P only or PA.

We also have two additional hypotheses:

3) Support at PA levels will yield the highest confidence in performing the judgment task compared to P only or PC.
4) Support at PA levels will yield the highest reliability across participants compared to P only or PC.

5.2.4. Apparatus to test judgment support hypotheses

An apparatus to investigate these hypotheses consisted of both judgment profiles with criterion values for each profile and a prototype tool to present the judgment profiles and record human judgments.

5.2.4.1. Judgment profiles

Thirty judgment profiles, or thirty sets of available cues specific to thirty different physicians, were created based on Cooksey's suggestion of a 10:1 profile to cue ratio. Profile creation consisted of two steps. Because data are managed in the EMR database at the patient-level, we had to first aggregate patient data to form populations of patients based on the primary care physician. We then had to derive the three available cues specific to each population.

We first obtained a report from the EMR database at UVa that included de-identified patient data that met the following inclusion criteria:

- Patients seen at the University Medical Associates (UMA) general medicine clinic within the last year
- Diagnosis of hypertension
- Between the ages of 18 and 74

The report contained the following for over 3,000 patients that met the criteria above:

1. Random number representing patient ID
2. Coded number to indicate primary care physician

3. Year of residency of provider – 1 (first year residency), 2 (second), 3 (third), 99 (attending physician)
4. Coded number to indicate resident firm (group of 5-6 residents assigned to a single attending physician) (if applicable)
5. Diagnosis of diabetes – 1 (for diabetes), 0 (no diabetes)
6. Blood pressure measure – 1 (≤135/85 mmHg for patients with diabetes and ≤140/90mmHg without), 0 (systolic or diastolic above these goals)
7. Recommended laboratory tests check – 1 (creatinine checked within last 15 months), 0 (creatinine not checked within 15 months)
8. Recommended medication – 1 (medication list includes ACE or ARB), 0 (medication list does not include ACE or ARB)

The patients were then grouped by primary care physician and the following metrics were calculated for each physician's panel population: percentage of patients at goal blood pressure, percentage of patients with diabetes on either ACE or ARB, and percentage of patients with creatinine checked within the last 15 months. These metrics represented the three available cues needed to judge the quality of hypertension care provided by that physician.

The thirty second and third year residents with the largest panel sizes (i.e., largest number of patients) were chosen as the thirty judgment profiles. The same three metrics were also calculated for three other populations of patients: the firm (specific to the resident, or profile, under consideration), all residents, and the entire clinic (which also included attending physicians' panels). Confidence intervals were then calculated for all populations. This was done using the Pearson-Klopper method for confidence intervals. However, to make the confidence intervals easier to interpret, we scaled the population percentages by a factor of ten. Without this scaling, the confidence intervals were so large that we thought the participants might disregard the data altogether. This scaling did not affect the displayed cues in any way as the displayed cues were already expressed as percentages.

Criterion values for each judgment profile were determined by averaging five internal medicine attending physicians' judgments for the thirty profiles. These attending physicians are considered subject matter experts in hypertension care. They had an average of 19.7 years of clinical experience (range 9-27 years) and an average of 17.8 years of academic medicine experience (range 9-24 years).

5.2.4.2. Prototype IA tool

In order to present the available cues, an IA prototype similar to the existing SPARC tool [23] was built. The tool was built in Microsoft PowerPoint and used Visual Basic and ActiveX controls to provide function for users to view the thirty judgment profiles at their own pace and to collect the user's judgments for each profile via a slider bar.

This prototype was instantiated to provide IA judgment support to the human at three level combinations based on our hypotheses presented above: "perceive" only, "perceive" plus "comprehend," and "perceive" plus "assess." At the "perceive" level, the IA transformed the three available cues to displayed cues using three dot plots, similar to how the current version of SPARC displays population-based data. Each dot plot consisted of three things: 1) the displayed cue represented by a black dot and a numerical indication of its value located at the bottom of the graph, 2) the confidence interval for each displayed cue represented by the width of a grey bar extending the height of the graph, and 3) other population values represented with orange dots, orange numerical values, and orange lines of length equal to that population's 95% confidence interval. An example dot plot is shown below.

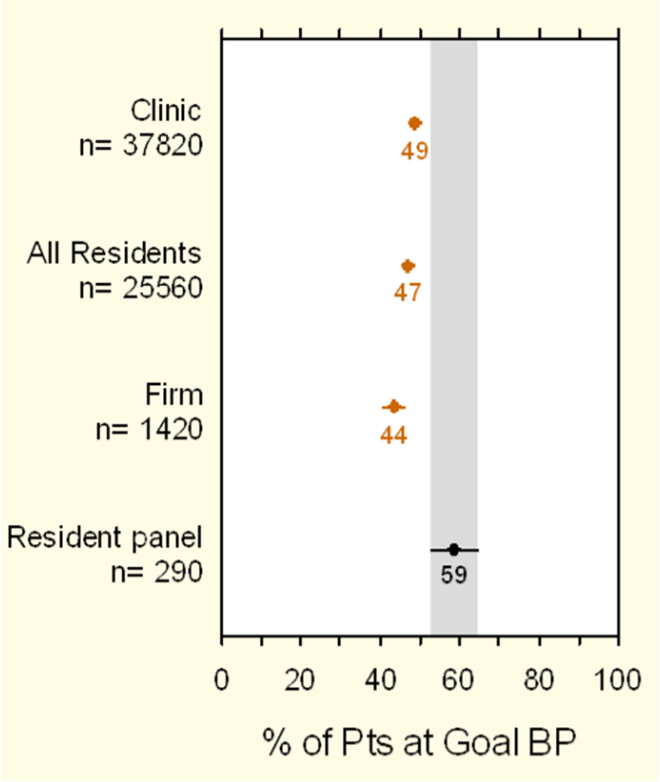
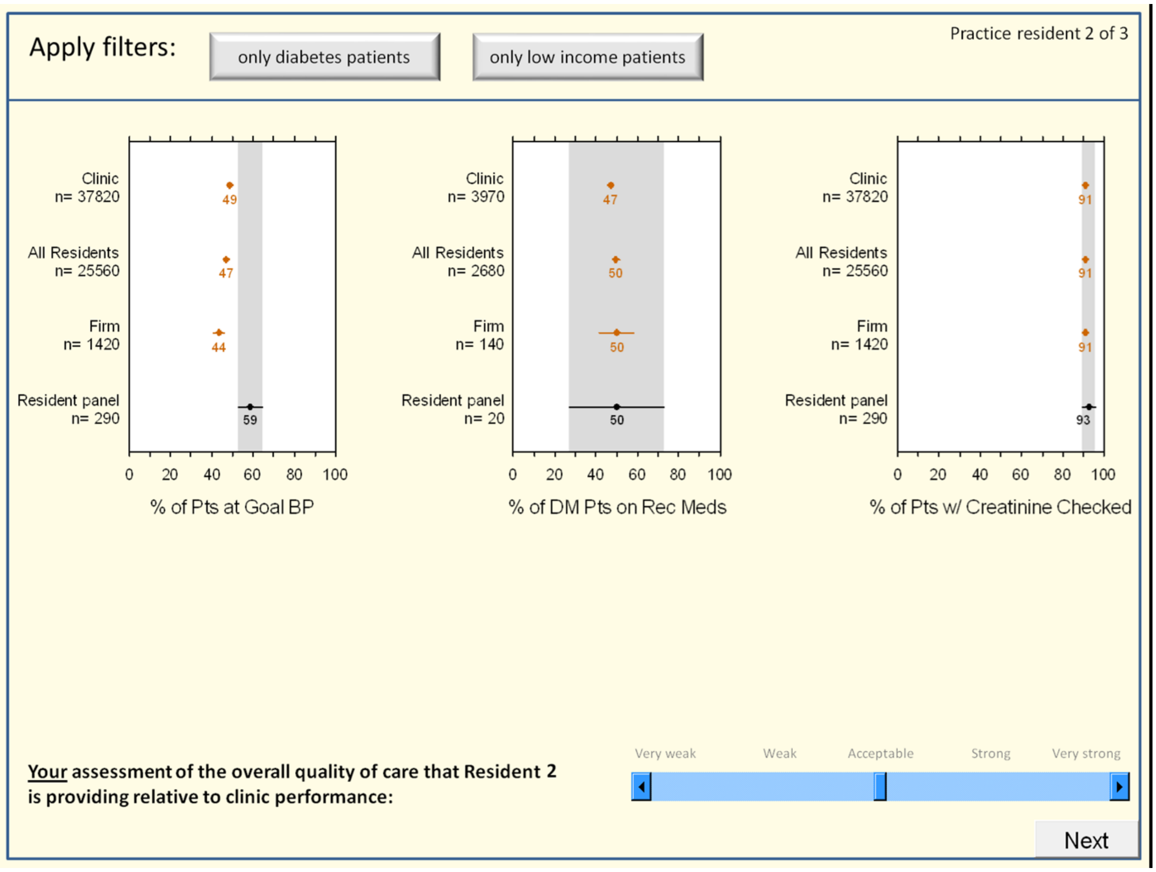


Figure 32. Dot plot used to represent displayed cues for health care quality judgments.

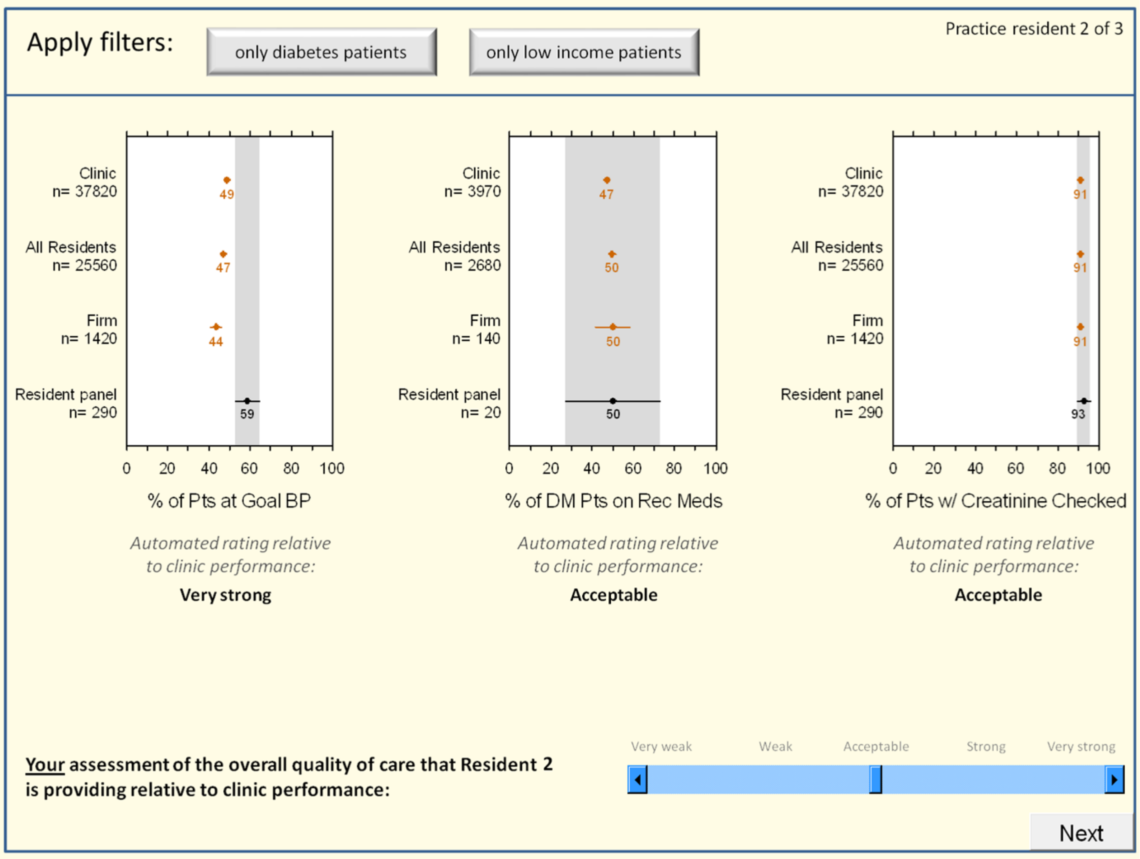
At the “perceive” level of judgment support, the displayed cues (dot plots) were presented simultaneously. The IA also provided filtering functionality similar to the current SPARC tool. Participants could view two filtered versions of the three displayed cues. One filter button changed the displayed cues from all patients with hypertension to patients with both hypertension and diabetes. The other filter button changed the displayed cues from all patients with hypertension to patients with both hypertension and low income. This instantiation of the IA prototype is shown in Figure 33.



Your assessment of the overall quality of care that Resident 2 is providing relative to clinic performance:

Figure 33. IA display for hypertension quality of care judgment task showing “perceive” (P) support level.

The IA prototype was also instantiated to provide judgment support at the “perceive” and “comprehend” levels. Added support to help comprehend the available cues involved providing an indication of the percentile rank for each available cue for that resident relative to all residents in the clinic. This percentile rank was expressed in one of five groups. The bottom percentile (0-20%) was represented with “very weak,” 20-40% was represented with “weak,” 40-60% was represented with “acceptable,” 60-80% was represented with “strong,” and 80-100% was represented with “very strong.” The percentile ranks were displayed below each dot plot. This version of the IA prototype is depicted in Figure 34.



Practice resident 2 of 3

Automated rating relative to clinic performance: **Very strong**

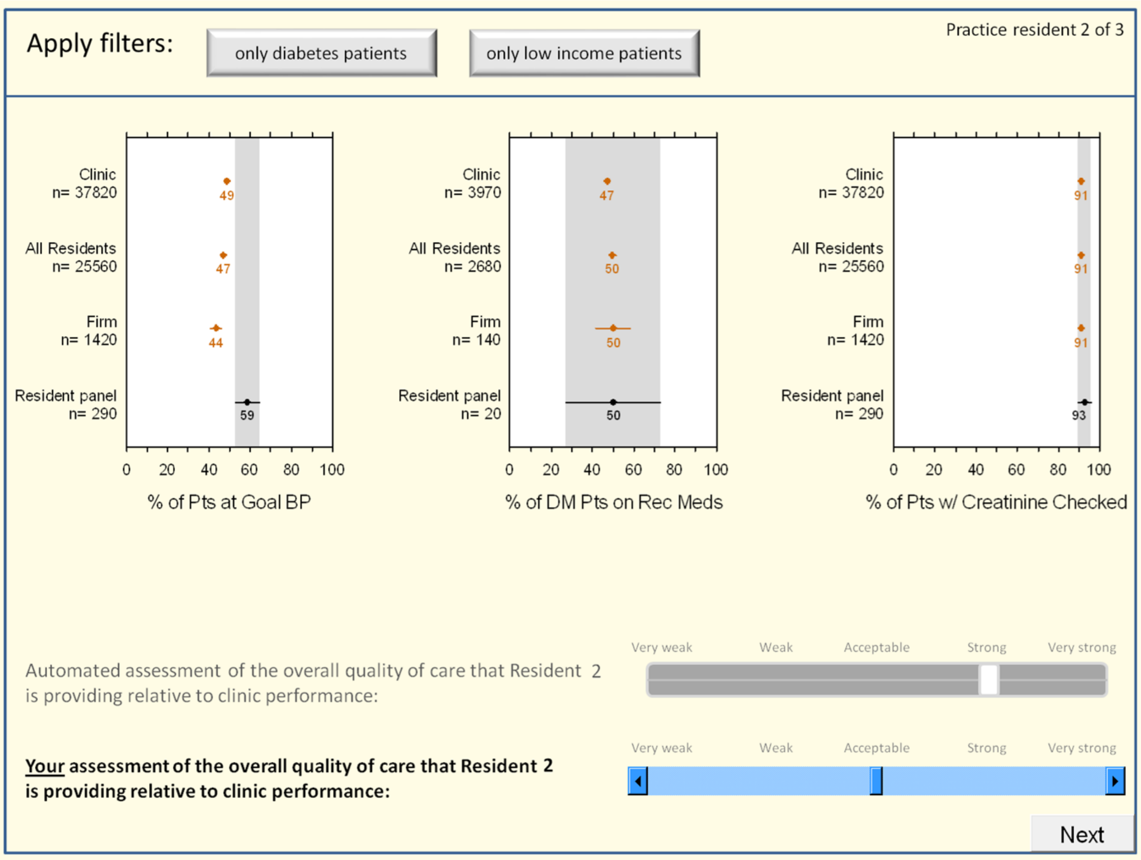
Automated rating relative to clinic performance: **Acceptable**

Automated rating relative to clinic performance: **Acceptable**

Your assessment of the overall quality of care that Resident 2 is providing relative to clinic performance: Acceptable

Figure 34. IA display for hypertension quality of care judgment task showing "perceive" and "comprehend" (PC) support levels.

The prototype was also instantiated to provide judgment support at the "perceive" and "assess" levels. For the "assess" level of support, the IA provided an overall quality of care score for the resident under consideration on a 5-point scale: "very weak," "weak," "acceptable," "strong," and "very strong". The strategy the IA employed to derive the overall quality of care score was to first calculate a weighted average for the three available cues for every resident in the clinic. The weighting scheme for this average was 2:1:1 and was determined by a focus group of six internal medicine attending physicians who agreed that the goal BP cue was twice as important as the other two cues. After calculating the weighted average for each resident, the resident was then grouped into percentile rank groups. The same percentile rank groups that were used at the "comprehend" level for each individual cue was used for the overall assessment. The percentile rank group was indicated with a tick mark at the appropriate percentile rank category. This version of the IA prototype is depicted in Figure 35.



Automated assessment of the overall quality of care that Resident 2 is providing relative to clinic performance:

Your assessment of the overall quality of care that Resident 2 is providing relative to clinic performance:

Figure 35. IA display for hypertension quality of care judgment task showing "perceive" and "assess" (PA) support levels.

5.2.5. Experimental design

5.2.5.1. Participants

Twenty-four resident physicians currently enrolled in the internal medicine program at UVa participated in the experiment. Participants ranged in age from 26 to 36 and included three females and twenty-one males. All participants were familiar with population-based patient data in that they all had previously participated in at least six, one-hour SPARC-related seminars where they were asked to investigate their panel of patients by viewing population-based dot plots similar to the ones used in this study.

5.2.5.2. Procedure used in data collection

Participants were first provided with a brief introduction to the study. They then stepped through a self-paced training session explaining how to use the automation's interface. This training also included three judgment practice trials to allow participants to gain familiarity with the task of judging quality of hypertension care provided by resident physicians.

During the experimental session, participants made thirty judgments about the quality of hypertension care provided by thirty resident physicians. Following their thirty judgments, they were also asked to rate their confidence in performing the task. The IA software recorded all responses.

5.2.5.3. Independent variables

Participants were grouped into one of three conditions of IA support for the duration of their thirty judgments. These groups are described in the following table and are correlated to the three instantiations of the prototype IA tool described in Figure 33 - Figure 35.

Table 10. Conditions of IA support.

IA level combination	**Description**
P	Displayed cues represented with dot plots
PC	Displayed cues represented with dot plots plus the percentile rank group for each displayed cue
PA	Displayed cues represented with dot plots plus an automated assessment regarding the overall percentile rank based on a weighted average of the three displayed cues

We used a simple blocked randomization scheme where three participants were randomized at a time. Thus, each participant was equally likely to be assigned to one of the three groups and each group was equal in size. There are six ways that one can randomize three participants equally into three groups. We used statistical software to randomly pick one of the six combinations before assigning the first participant for each group. No participant characteristics were considered in treatment allocation.

5.2.5.4. Dependent variables

For every judgment profile, the following data were collected: the environmental criterion, available cues, displayed cues, and the human's judgment. An overall confidence in the judgment task was also collected for each participant.

From these data, derived measures based on the ELMA framework were calculated (see Table 5). The following ELMA measures were derived to characterize the context in which the human made judgments:

- Available environmental predictability, $R_{E.a}$

- Displayed environmental predictability, $R_{E.d}$
- Displayed accuracy of the automation, $V_{d.a}$

To characterize the human judge (and thus, the human-automation system) the following ELMA measures were calculated. (Note that we did not ask participants for their interpretation of the displayed cues, to collect subjective cues. Thus, we cannot calculate the human's subjective cognitive control, $R_{H.s}$ or the subjective accuracy of the human judge, $V_{d.s}$.)

- Human-automation judgment achievement, r_a
- Displayed cognitive control, $R_{H.d}$
- Displayed linear knowledge, $G_{H.d}$

5.2.5.5. Data analysis

Four main steps were followed using the ELMA framework. First, the environmental context was characterized by modeling the criterion using the available and displayed cues and then computing the relevant ELMA measures. Second, linear models were created for each participant using the displayed cues and then computing the relevant ELMA measures to characterize their judgment performance. Third, nomothetic analysis of the ELMA measures for IA support effects was conducted. Last, post-hoc analysis of variance in judgment performance between the best and worst participants (in terms of human-automation judgment achievement) was conducted.

The experimental design for nomothetic analysis was a fixed effects model design, with IA support condition as the fixed effect. Multiple Analysis of Variance (MANOVA) was used to test for effect of level of support on the ELMA measures with adjusted alpha levels. Univariate ANOVAs were subsequently conducted for all significantly impacted ELMA measures. Post hoc analysis was also conducted using Tukey's Honestly Significant Difference (HSD).

All derived measures from the ELMA framework are correlations. Therefore, before performing the nomothetic data analysis described above, the correlations were transformed using Fisher's r to z_r transformation (equation 4.2) to obtain normally distributed variables as suggested by Cooksey [40]. However, descriptive statistics of the untransformed measures are reported and graphed.

5.3. Results

The results of this experiment are presented using α = 0.05 for significance and α = 0.1 for a trend. The environmental context is discussed first. This is followed by a discussion of the nomothetic analyses investigating the impact of IA support on human-automation judgment performance. An idiographic investigation of judgment performance of two participants is then discussed.

5.3.1. Environmental context

To characterize the context in which the human must make judgments, we can start by modeling the environmental criterion using both available cues and displayed cues. This multiple regression modeling results in the following models where a_i and d_i are the available and displayed cues respectively of percentage of patients at goal blood pressure, percentage of diabetes patients on recommended medications, and percentage of patients with their creatinine checked within the last 15 months.

$$quality\ of\ care = -154.04 + 2.06(a_1) + 0.63(a_2) + 0.80(a_3) \tag{5.1}$$

$$quality\ of\ care = -155.51 + 2.04(d_1) + 0.63(d_2) + 0.83(d_3) \tag{5.2}$$

Both models of the criterion were checked to ensure proper assumptions of normality, linearity, homoscedasticity, and independence of residuals. Also, both sets of available cues and displayed cues are expressed as percentages of patients, so the cue values would not influence the regression weights in the models. All parameters in both models were found to be significant at $p < 0.001$, indicating that all three cues have a significant effect on the criterion of quality of care.

These models indicate that the first cue of percentage of patients at goal blood pressure has the greatest impact on the quality of hypertension care provided by a physician (more than twice the impact that the other cues have). This relative impact of cues on quality of care could be indicative of the clinical focus of the attending physicians in that they are more concerned with patient outcomes (maintaining goal blood pressure) compared to process measures (patients on recommended medications and with recommended laboratory tests checked).

From the models of the criterion, we are able to derive the available and displayed environmental predictability for the judgment task in this experiment. These measures indicate how predictable the criterion is based on the cues, which essentially represents an upper bound on the judgment performance of the human judge.

Correlating the environmental criterion and the model of the criterion based on available cues resulted in an available environmental predictability of 0.97. Similarly, correlating the environmental criterion and the model of the criterion based on displayed cues resulted in a displayed environmental predictability of 0.97. Thus, the displayed accuracy, $\mathrm{V}_{\mathrm{d.a}}$, of the automation is equal to 1, which is a characterization of the transformation of available to displayed cues.

$$R_{E.a} = cor(E, M_{E.\mathrm{a}}) = 0.97 \tag{5.3}$$

$$R_{E.d} = cor(E, M_{E.d}) = 0.97 \tag{5.4}$$

$$V_{d.a} = \frac{R_{E.d}}{R_{E.a}} = 1 \tag{5.5}$$

5.3.2. Nomothetic results of judgment performance

The impact of IA support on human-automation judgment performance across groups of participants is presented in this section, beginning with an examination of the participant judgment models using the displayed cues. This is followed by an analysis of the ELMA measures and an investigation into the impact of IA support on reliability of judgments across participant groups. Results related to confidence in judgment are also presented.

5.3.2.1. Examination of participant linear models

The first step in examining judgment performance of the participants is to create one linear model for each participant. Figure 36 indicates the computed cue weights for all participant models. The criterion cue weights (based on the model using displayed cues) are also depicted with the star and dotted line. The extent to which the participants' cue weights mirror the cue weights of the criterion model is indicative of how well participants used the displayed cues to make their quality of care judgments. The first and second cues tended to have less of an effect on the quality of care judgments for the residents compared to their effect on the criterion. There is also more variance in the cue weights for the first and third cues compared to the second cue, indicating that across participants the second cue had a more consistent degree of effect on the residents' judgments. In general, the patient outcome measure (percent of patients at goal BP) had less of an effect on resident judgment of quality of care compared to attending judgments (criterion). Further, the process measure of percent of patients with recommended laboratory tests checked had a greater impact on five residents' judgments of quality of care compared to the criterion. This could indicate that such residents may be more concerned with process measures, which they have more control over and less concerned with outcome measures that may be difficult to control in short periods of time (i.e., their residency program).

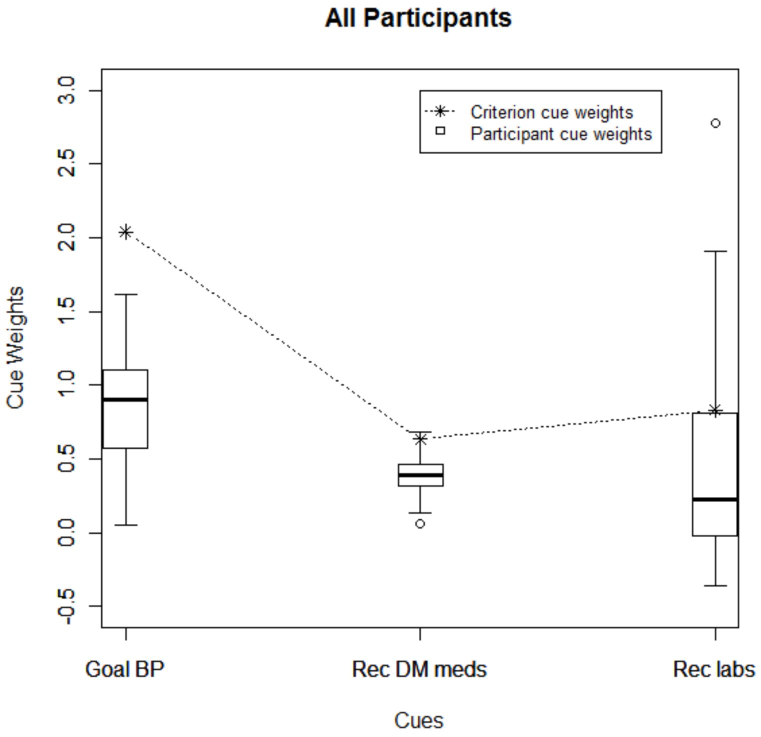


Figure 36.Cue weights for all participant models.

To further explore the participant models, we can examine cue weights for each group of IA support independently. These cue weights are shown in the following figures. Observations of these plots indicate the participants in the PC group may have used the displayed cues more closely related to the criterion's model. However, the PC group had more variance in the effect of the third cue compared to the other two participant groups. For participants in the P group, all cues had less impact on judgment compared to the criterion model, particularly for the first cue of percentage of patients at goal blood pressure.

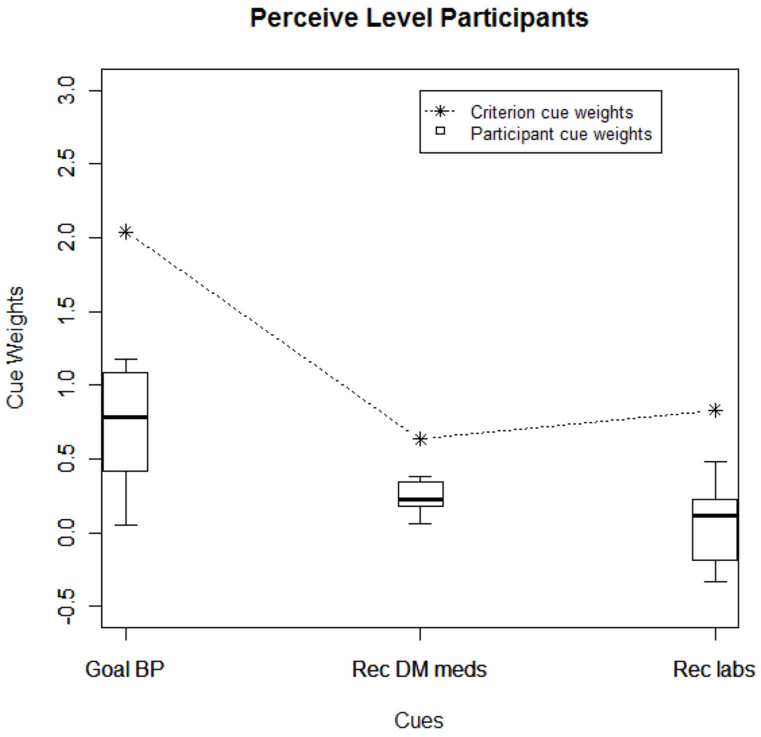


Figure 37. Cue weights for participants in the perceive (P) condition of IA judgment support.

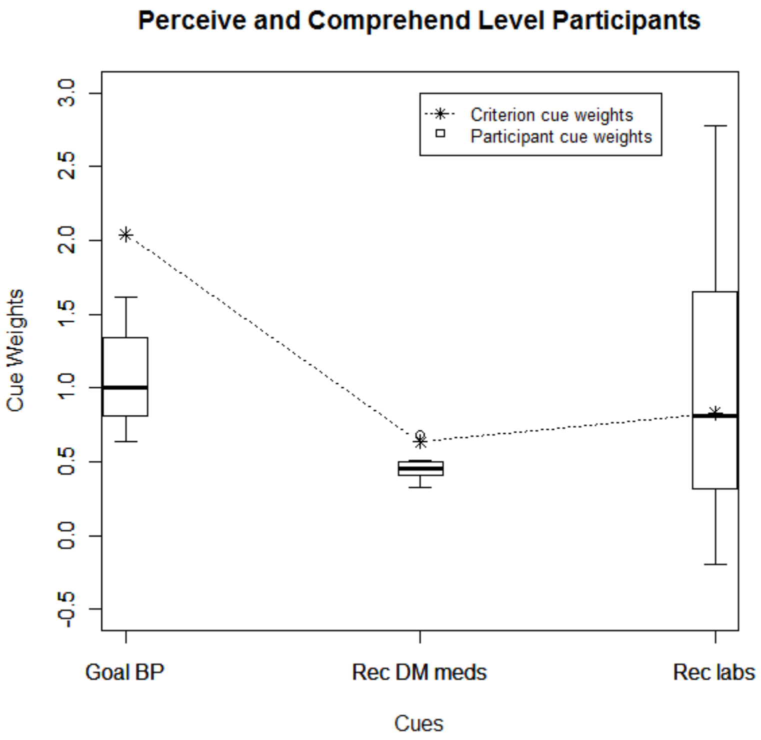


Figure 38. Cue weights for participants in the perceive and comprehend (PC) condition of IA judgment support.

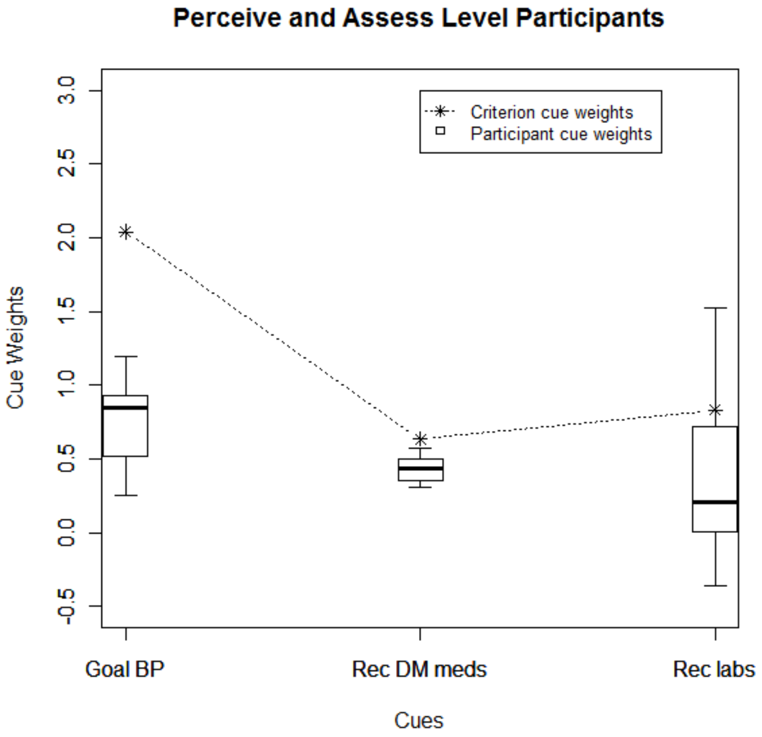


Figure 39. Cue weights for participants in the perceive and assess (PA) condition of IA judgment support.

5.3.2.2. ELMA measures

Given the similarities and inconsistent conclusions based on observing the models, it is still not clear how the amount of IA support impacted judgment performance. Thus, the next step of analysis is to compute the ELMA measures to characterize each participant's judgment performance using the criterion-participant model pair as illustrated in Figure 29, Figure 30, and Figure 31. The predicted criterion and predicted human judgment values, the actual criterion and human judgment values, and the residuals were obtained from the regression analysis and used to derive the ELMA measures of human-automation judgment achievement (r_a), displayed cognitive control ($R_{H.d}$), displayed linear knowledge ($G_{H.d}$), and displayed un-modeled agreement ($C_{H.d}$).

The average for each ELMA measure was calculated for all participants and for each of the three groups of participants by condition of IA support. These results are shown in Figure 40. The ELMA measures from left to right are achievement, displayed cognitive control, displayed linear knowledge, displayed un-modeled agreement, and displayed environmental predictability. From this graph, we can see that the PC condition of IA support yielded better human-automation judgment achievement and displayed cognitive control. This is consistent with the above investigation of cue weights where it appeared that the PC group more closely matched the cue weights of the criterion model.

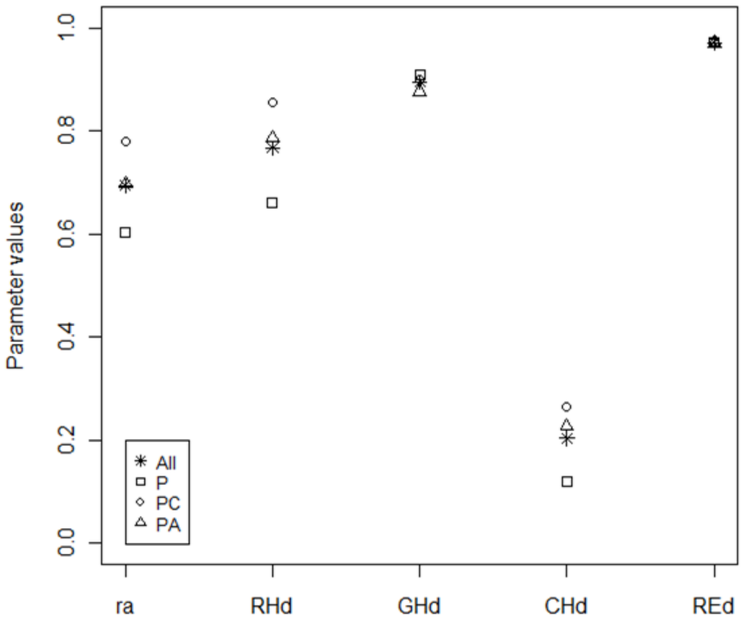


Figure 40. Average ELMA measures for all participants and for each IA support group of participants.

Considering all of the ELMA measures, except R_{Ed} that did not vary between IA support conditions, there is a statistically significant impact of IA support on judgment performance overall (F (8, 36) = 4.05; p < .005; Wilk's λ = 0.277). (To confirm appropriate assumptions, Levene's Test of equality of error variances showed that all measures had homogeneity of variances (p < .05).)

Subsequent univariate analyses showed that condition of IA support had a statistically significant effect on both human-automation judgment achievement (r_a) (F (2, 21) = 7.05; p < .005) and displayed cognitive control ($R_{H.d}$) (F (2, 21) = 9.61; p < .002) (the first two ELMA measures represented in Figure 40. All ANOVA results are presented in the following table.

Table 11. ANOVA results of ELMA parameters across IA support condition.

Effect	Human-automation judgment achievement, r_a	Displayed cognitive control, $R_{H.d}$	Displayed linear knowledge, $G_{H.d}$	Un-modeled agreement, $C_{H.d}$
IA support	F (2, 21) = 6.52 p < .007 *	F (2, 21) = 9.61 p < .002 *	F (2, 21) = 0.60 p = 0.56	F (2, 21) = 1.56 p = 0.23

Figure 41 depicts human-automation achievement, r_a, for each IA support condition in more detail. Tukey's post hoc analysis indicates that the mean of human-automation achievement was significantly higher for the PC condition compared to the P condition (p < .005).

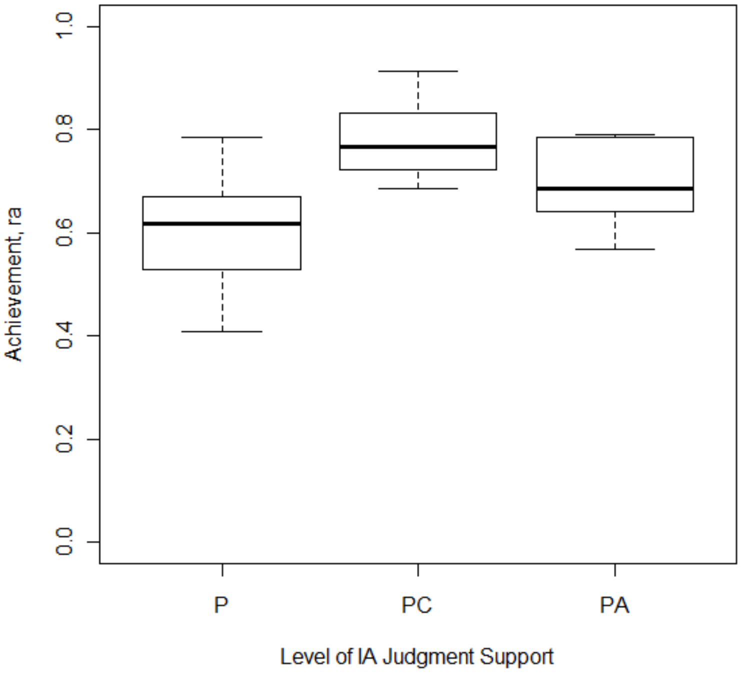


Figure 41. Human-automation judgment achievement by IA support condition.

Based on the ELMA lens model (equation 3.23), we know that human-automation judgment achievement is dependent on displayed cognitive control, displayed linear knowledge, displayed environmental predictability, and un-modeled agreement. Displayed environmental predictability ($R_{E.d}$) did not vary across participants in this experiment. Thus, we can investigate the other ELMA measures in more detail to further understand the differences in judgment achievement between the IA support conditions.

Figure 42 depicts displayed cognitive control, $R_{H.d}$, for each IA support condition in more detail. The graph also includes a dotted line representing the displayed environmental predictability, $R_{E.d}$. The PC group exhibited displayed cognitive control closest to the environmental predictability. This implies that the PC group was as consistent in applying their models almost as much as the criterion was linearly predictable, while the P and PA groups were less consistent. Tukey's post hoc analysis indicates that the mean of displayed cognitive control was significantly higher for the PC condition compared to the P condition (p < 0.001). There were also trends for the PC condition to be higher than the PA condition (p = 0.1) and the PA condition to be higher than the P condition (p = .08).

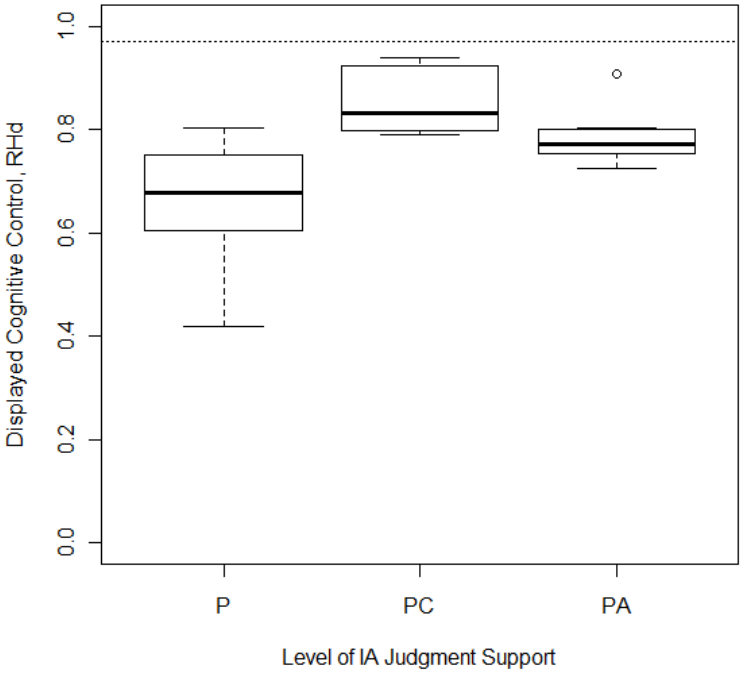


Figure 42. Displayed cognitive control by IA support condition.

Human-automation judgment achievement and displayed cognitive control appear to improve simultaneously as IA support increases from the P to PC conditions. There is some improvement from P to PA conditions, although it seems that judgment achievement and displayed cognitive control actually decrease as the amount of IA support increases from PC to PA.

One hypothesis for this decrease in judgment performance when participants are provided with support at the "assess" level is that the participants with the automated assessment are correlating their judgments to the IA's judgments, which may be different than the criterion. However, the following plot shows that this is not the case. Participants in the PC condition are actually more correlated with the automation's assessment compared to the PA condition, even though the PC group did not see the automation's assessment during any part of the experiment. Further, the dotted line on the graph shows the correlation between the criterion and the automation to be 0.92. So even if participants were more correlated with the automation, this should not result in a decrease in judgment achievement because the automated judgment was more correlated than even the best participant (in any IA condition) with the criterion.

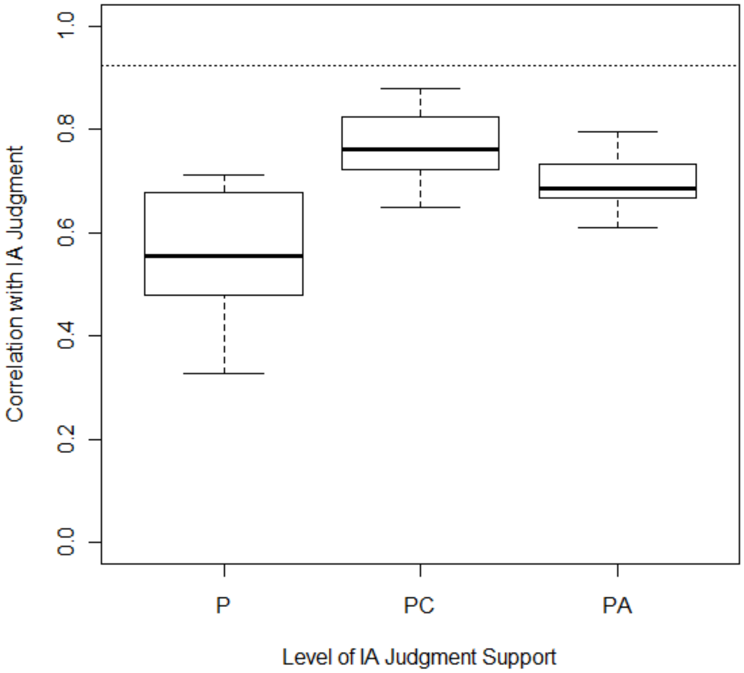


Figure 43. Participant judgments correlated with IA judgment across conditions of IA support.

There were no significant differences in displayed linear knowledge or un-modeled agreement across the groups of IA support. This indicates that participants in each group could be modeled with policies that accurately reflected the linear relationships between the criterion and the displayed cues. Boxplots depicting these measures are shown in Figure 44 and Figure 45. Thus the difference in human-automation judgment achievement based on the decomposition of the ELMA lens model equation can primarily be attributed to participants' differences in cognitive control in that with IA support at the "comprehend" level they were able to more consistently apply their judgment policies and were better correlated with the criterion.

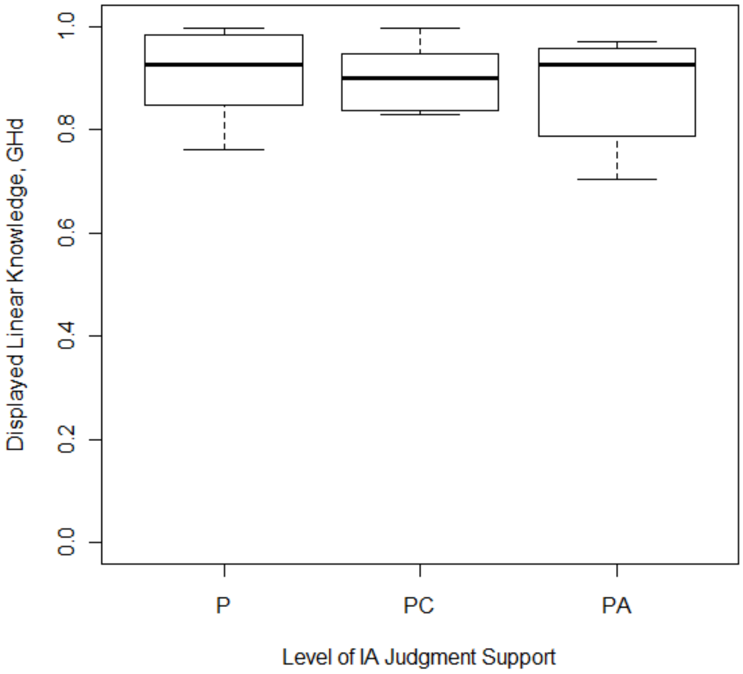


Figure 44. Displayed linear knowledge by IA support condition.

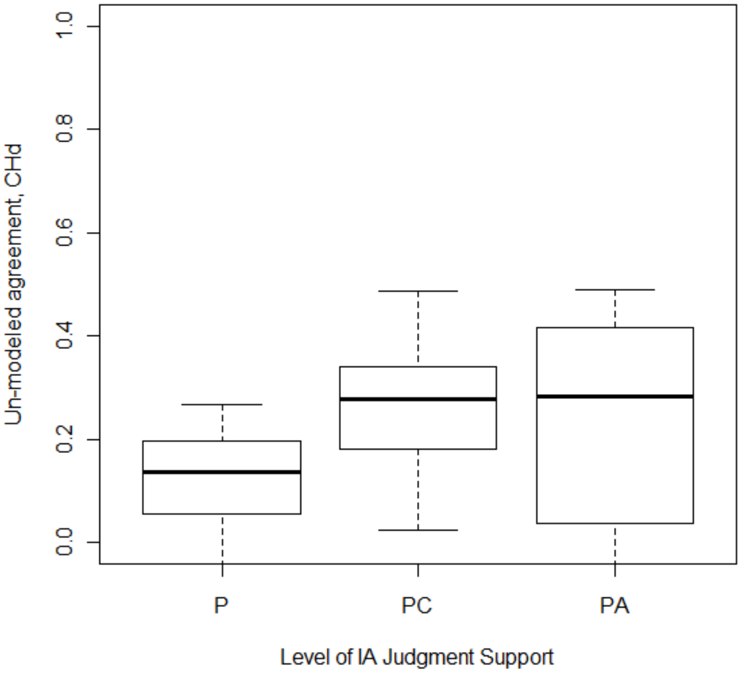


Figure 45. Un-modeled agreement by IA support condition.

5.3.2.3. Reliability across participants within IA support conditions

The amount of IA support also had an impact on reliability across participants. Interestingly, reliability was found to be significantly better at a lower level of IA judgment support. To measure reliability across participants as a function of IA judgment support condition, we investigated the standard deviation for each judgment profile across participants within IA support condition. The following plot shows that the P condition of judgment support results in the smallest average standard deviation of participant judgments. A repeated measures ANOVA with IA support condition nested within profile (i.e., we can see the “IA support” effect within each and every “judgment profile”) shows that IA support significantly impacts the standard deviation of the judgment profiles across participants. Post-hoc pair-wise t-tests with adjusted p-values show that the PA ($\mu = 17.2$) condition results in significantly higher standard deviations compared to both the P ($\mu = 10.2$) ($p < 0.0005$) and the PC ($\mu = 11.2$) ($p < 0.0005$) conditions.

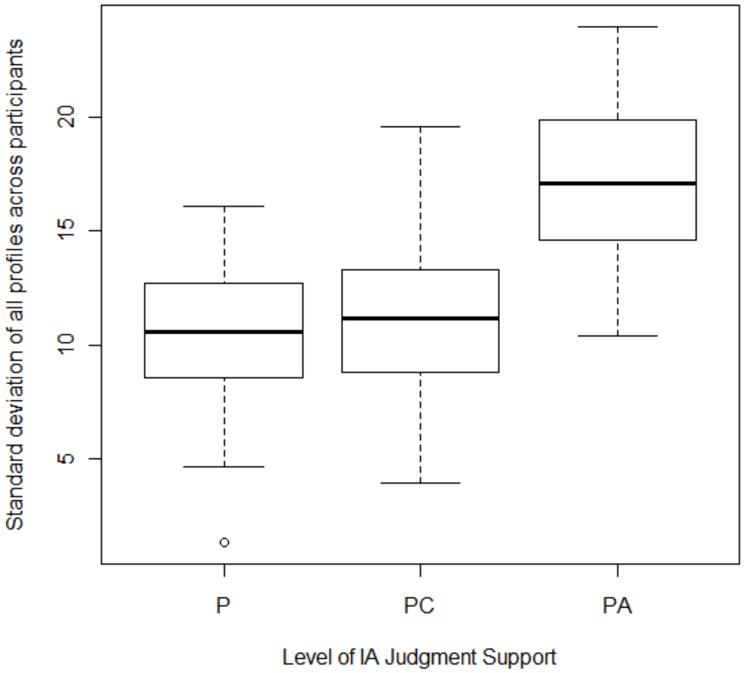


Figure 46. Standard deviation of all profiles across participants grouped by IA support condition.

One hypothesis as to why the PA condition appears to impact judgment performance in terms of increased standard deviation (reduced reliability) across participants is that dissonance between the automated assessment and the displayed cue of percentage of patients at goal blood pressure (the cue with the highest average weight ($\beta_1 = 0.85$) caused the participants to distrust the automated assessment and then bias their judgments further from the criterion (and the automation's assessment). To examine this hypothesis, we first split the judgment profiles into two groups: 10 profiles where the automated assessment deviated by more than two categories that the "comprehend" level of support would have shown the participants for that cue had they been provided with that level of support (they were not) and 20 profiles where the automation assessment deviated by one or zero categories of "comprehend" level support for that cue. We then recalculated human-automation judgment achievement for participants in the PA group under both conditions (dissonance and no dissonance). However, we found that only two of the eight participants had better judgment achievement in conditions of no dissonance compared to conditions of dissonance. We also calculated Murphy's skill score [71] for participants under both conditions and found that four out of eight participants had better skill scores under no dissonance. Thus, this specific bias does not appear to be a contributor to poor reliability across the PA group of participants for this task.

Another way to investigate reliability across participants is to correlate all pairs of judges in each IA support condition (28 pairs of 8 judges in each of the three IA conditions). The average correlation in the P group of participants was 0.52, the average in the PC group was 0.67, and the average in the PA group was 0.64. This analysis does not yield the results that the PA group was worse than the P group in terms of reliability; however, this provides more evidence that providing residents with IA support at the "comprehend" level may help judgment performance in terms of increasing reliability among judges.

5.3.2.4. Confidence in judgment performance

Overall, participants were not very confident in their ability to judge the quality of hypertension care provided by the residents in this study ($\mu = 44/100$, $\sigma = 19$). Despite the differences in the amount of IA support between participant groups, we found no significant effect on participant's confidence in their judgments. This could be due to participants having not yet spent much time with the tool. These results are shown in Figure 47.

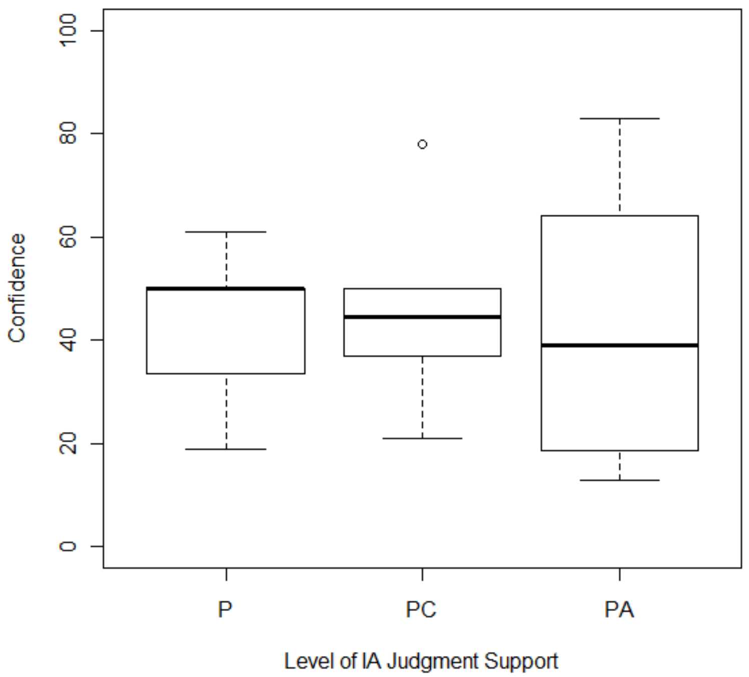


Figure 47. Confidence of participants in their judgments by IA support condition.

5.3.3. Idiographic results of judgment performance

In addition to investigating judgment performance across participants grouped into IA support conditions, the highest and lowest achieving (based on human-automation judgment achievement) participants were explored in greater detail. The following analysis demonstrates how the ELMA framework can provide insight into individual judges. The lowest achieving participant was assigned to the P group of IA support and had judgment achievement equal to 0.41. The highest achieving participant was assigned to the PC group of IA support and had judgment achievement equal to 0.91.

First it is useful to compare the cue weights between the two participant models and the cue weights used in the criterion model based on the displayed cues. Figure 48 and Figure 49 show how the cue weights of the two participant models compared to the criterion model. From these figures we can see that the highest achieving participant used cue weights that more closely matched those of the criterion model. Further, the cues had less impact on the lowest achieving participant compared to the impact the cues had on the criterion. The intercept values for both participant models were also drastically different (β_0 = -113 for the highest achieving participant; β_0 = 27 for the lowest achieving participant; β_0 = -133 for the criterion model). However, the first cue of percentage of patients at goal blood pressure had the largest effect on both participants' judgments. It is important to note that neither participant was told the relative importance of the cues.

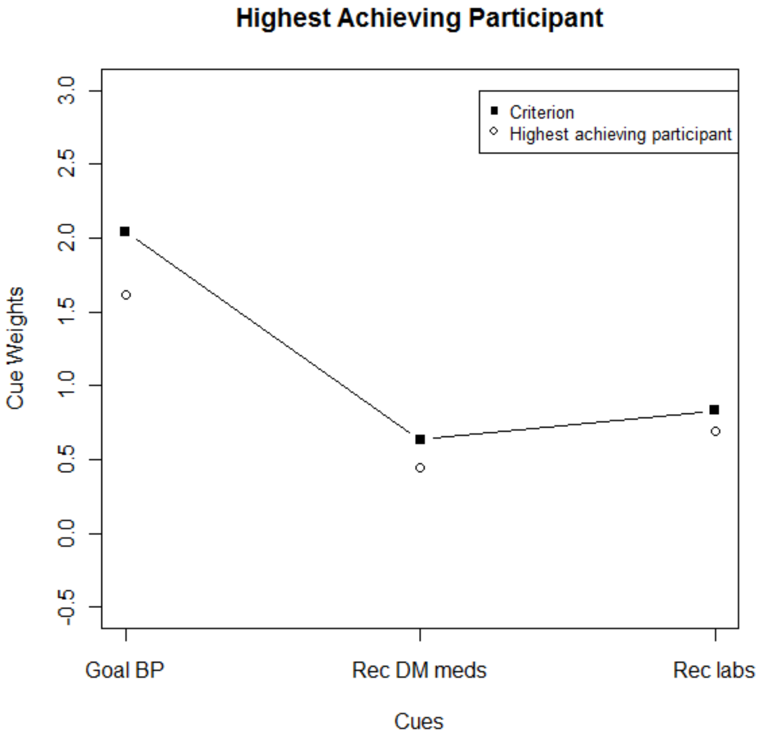


Figure 48. Criterion and highest achieving participant cue weights.

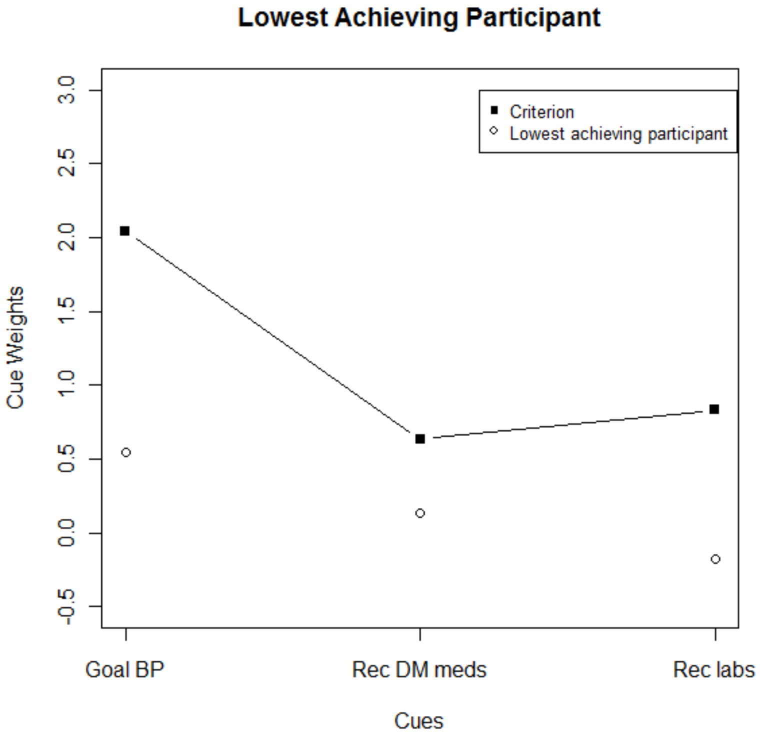


Figure 49. Criterion and lowest achieving participant cue weights.

From the ELMA lens model equation based on displayed cues (equation 3.32), human-automation judgment achievement (r_a) is dependent on displayed cognitive control ($R_{H.d}$), displayed linear knowledge ($G_{H.d}$), un-modeled agreement ($C_{H.d}$), and displayed environmental predictability ($R_{E.d}$). We know both participants made judgments under the same environmental predictability. However, we can investigate the other ELMA parameters to uncover differences in judgment achievement. Figure 50 depicts the ELMA parameters for the two participants and the average of all participants.

The two participants greatly differed in their displayed cognitive control ($R_{H.d}$ = 0.91 vs. 0.42). The highest achieving participant exhibited cognitive control close to the displayed environmental predictability ($R_{E.d}$ = 0.97). This indicates that this participant's judgments were as consistent with a linear model as the environment was linearly predictable, while the other participant was much less consistent. The higher achieving participant also exhibited better displayed linear knowledge ($G_{H.d}$). This indicates that the higher achieving participant could be modeled with a judgment policy which accurately reflected the linear relationship between the cues and the criterion. Participants had similar values of un-modeled agreement ($C_{H.d}$), indicating that both were appropriately influenced by nonlinear relationships between cue values and the criterion.

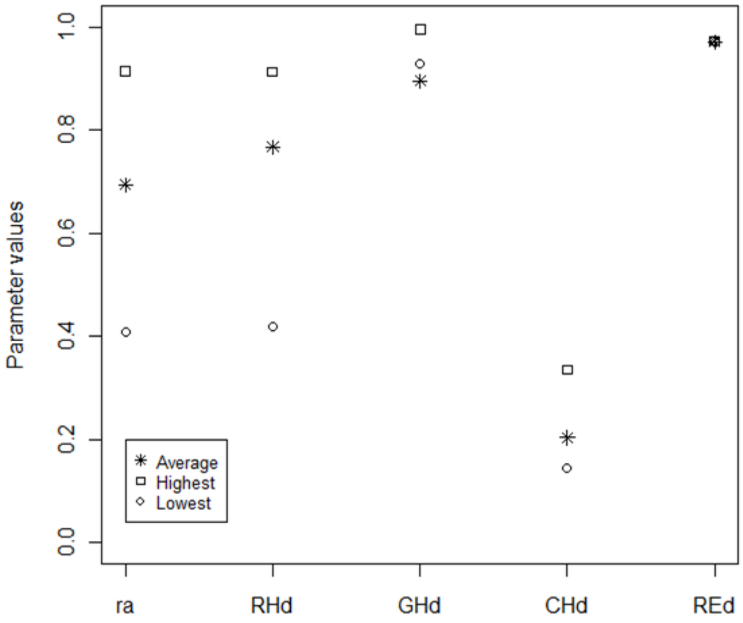


Figure 50. ELMA measures for the highest and lowest achieving participants and the average of all participants.

Thus, the ELMA lens model analysis indicated that differences in human-automation judgment achievement in this task can be attributed primarily to differences in the participants' ability to consistently apply their judgment policies. The structure of their policies also contributed to the differences in judgment achievement.

Despite having the highest human-automation judgment achievement, the participant reported their confidence in performing quality of hypertension care judgments ($conf_{high}$ = 21) to be more than one standard deviation below the average confidence ratings across all participants (μ = 43.7, σ = 19.1). The lowest achieving participant reported their confidence to be greater than the average confidence rating ($conf_{low}$ = 50).

5.4. Discussion

This was the second instantiation of the ELMA framework. We used ELMA to understand the judgment task of assessing population-based quality of hypertension care. ELMA guided the identification of the available cues needed for the judgment task and helped to conceptualize the task in order to derive hypotheses regarding the effect of automation support on judgment performance. ELMA also guided the analysis of the environmental context and human-automation performance at both a nomothetic and idiographic level. These analyses demonstrated the usefulness of the ELMA framework, particularly to investigate the effects of varying conditions of IA judgment support on judgment performance and to investigate sources of individual differences in judgment performance. The results have implications for both the design of IA tools and training interventions targeting specific aspects of the human-automation judgment process.

5.4.1. Effect of automated support on human-automation judgment performance

One main objective of this research was to investigate the effect of automation support on judgment performance. Resident physicians provided with IA support at the “perceive” and “comprehend” (PC) levels, had significantly better human-automation judgment achievement and displayed cognitive control compared to residents provided with only “perceive” IA support. Further, there was a trend for better cognitive control for residents provided with PC support compared to those provided with “perceive” and “assess” support. This partially confirms our first and second automation design hypotheses that PC support would yield better judgment achievement and cognitive control compared to P or PA conditions. At the individual judge level, we also found that differences in judgment achievement were predominantly due to differences in participants’ ability to consistently apply their judgment policies. This could be a result of there being no generally accepted standard for quality of hypertension care, including a general acceptance for the relative importance of each cue.

Similar results were found in the first application of the ELMA framework presented in Chapter 4 of this dissertation. In both studies, added support at the comprehend level may have contributed to enhanced cognitive control in that participants were able to better understand the environmental context (i.e., the cues) and were thus able to more consistently apply their judgment strategies, resulting in greater judgment achievement. Similar results were also found by Bisantz et al. [19] and by Strauss and Kirlik [9] when they determined that individual differences in judgment achievement were most attributed to differences in participants’ ability to consistently apply their judgment strategies, rather than differences in task knowledge.

Confidence in ability to judge the quality of hypertension care was not affected by the condition of IA support. In general, participants rated their confidence fairly low no matter what amount of IA support they had to make their judgments. This is most likely due to the judgment task itself. We know that physicians are not good at making self-assessments regarding quality of care [16], which may be similar to rating other peer physicians. Further, many of the attending physicians we interacted with throughout this study expressed their discomfort in assigning a resident a quality of care score.

The amount of IA support did have a significant impact on the reliability of judgment performance across participants. Participants provided with PA support exhibited significantly less reliability in their judgments compared to the other two conditions of judgment support. This result is not in agreement with our fourth hypothesis. In fact, we found the opposite effect in that PA support resulted in reduced reliability across participants. This could have significant implications for judgment contexts in which multiple judges are expected to make similar assessments when given the same cue context. For example, insurance companies may use quality of care judgments made by different judges to make decisions regarding reimbursements. In this context, reliability across judges would be essential. It is unclear why the PA condition hindered reliability of residents' judgments and future work should address this issue.

One limitation to this study was that attending physicians' judgments were used as the criterion for quality of care. Thus, the criterion values could have been biased to the focus of the clinic in which they work or to the individual pressure or interest they each have related to the three available cues. For example, it could be that the attending physicians are more interested in maintaining goal blood pressure among their hypertensive patients because that outcome is under more scrutiny by insurance companies or hospital administrators. These biases could have impacted the ELMA measures of human-automation judgment achievement and displayed linear knowledge. Despite this limitation, we were still able to uncover effects of IA support on displayed cognitive control and reliability across participants, which were measured independent of the criterion.

5.4.2. Implications for the design of automation

These results have implications for the design of automation to support human judgment. For this judgment task, the true cues are unknown and impossible to practically measure. Thus, the displayed cues' uncertainty is tied directly to the uncertainty between the true cues and the available cues in the environment. Although providing residents with an automated judgment of the quality of hypertension care may improve judgment achievement compare to those provided with only the cues (at the "perceive" level), the added "assess" support may decrease the reliability across resident judges or the consistency within an individual judge. Thus, it may be more beneficial to focus design efforts on providing “comprehend” level support, or support on how to interpret and understand the available cues needed to judge quality of care. Residents provided with an IA prototype of this level of support were able to make quality of care judgments similar to those of attending physicians. The residents were also able to more consistently apply their judgment strategies to the level of predictability of the attending judgments.

Because differences in judgment achievement were due to participants’ inability to consistently execute judgment strategies rather than their knowledge of the task environment, automation design and training should potentially be focused on how to support consistent execution of judgment strategies. This could be implemented in the form of cognitive feedback based on the lens model framework. Balzer et al. [105] suggest that task information feedback during training could provide the greatest improvement on human judgment performance. In the cognitive feedback context, task information includes the environmental predictability, cues-criterion relationships, and inter-cue relationships. However, providing participants with cognitive information (measures of their own cognitive control and cue-judgment relationships) during a medical diagnosis task improved overall cognitive control [115]. The authors concluded that this information enabled participants to understand their own judgment policies, and thus make judgments more consistently.

There are numerous ways to represent the displayed cues and the support of the automation. Thus, in order to fully generalize the results found here, further research using different representations should be conducted. It may also be worthwhile to investigate the impact of display representations on the interpretation of the displayed cues (in the form of subjective cues). This issue was outside of the scope of this dissertation; however, the ELMA framework provides a platform to investigate the displayed to subjective cue transformation in different judgment contexts. We also found no differences in the confidence residents had in judging quality of care. It may be useful to explore different display representations that may increase confidence.

From an application perspective, it is important to understand the impact that automation design has on quality of care judgments. However, it is unclear if clinicians who are better able to judge the quality of care will actually improve their practice behaviors, resulting in better outcomes for the patients. Thus, it would be necessary to investigate any effects of either the judgment process or outcomes of the judgment process on practice behaviors, particularly with resident physicians who must demonstrate their ability to investigate and evaluate their practice.

6. Conclusions and future directions

This dissertation presented the Expanded Lens Model with Automation (ELMA) framework. ELMA is a useful tool for systems engineers as it provides a systematic framework to inform automation design choices and a quantitative method to evaluate human-automation judgment systems. ELMA accounts for discrepancies between how cues in the environment are transformed into displays to operators via automated processes. The transformation is based upon the desired, hierarchical level of cognitive judgment support. ELMA also includes quantitative measures to evaluate the human-automation system with an idiographic-statistical approach.

Two judgment tasks were investigated to demonstrate the utility of ELMA. Across both tasks, ELMA revealed that automated cue comprehension support improved judgment achievement. ELMA also revealed that the differences in achievement were predominantly due to the consistency with which participants used the displayed cues to make their judgments. This has implications for potential training interventions and may be further explored with different display representations. Results also suggest that reliability may be affected by providing automated assessment support in quality of health care judgment tasks. This could impact contexts where multiple judges are expected to make the similar judgments using the same cue sets.

However, it is challenging to disentangle the effects of the level of automated judgment support (i.e., the functionality of the automation) resulting in the content of information (via displayed cues) and the specific representations of the information (i.e., the display design). Thus, to draw stronger conclusions regarding the impact of level of automated support on different judgment tasks, additional display representations must be investigated. ELMA provides a systematic method for this investigation and can be used to guide iterative automation design choices.

Another limitation of ELMA is that it requires the judgment task to be defined and analyzed in accordance with the structure of the ELMA lens model. The task must be defined as a known (or estimated) criterion to be judged, based on a set of cues. This does not comprise the task of selecting among decision alternatives that may result from the judgment task, which in some contexts may be of more importance than the judgment. Further, there may be tasks where decomposing judgment from decision making is difficult. However, the ELMA framework could be used in conjunction with the PSW model of types and levels of automation that does account for both judgment and decision making [1] or with other decision modeling and analysis techniques, such as rule-based models [131], fuzzy rule-based models [132], operator function models [133], or naturalistic decision making methods [134].

Future extensions of ELMA could also include scaling the ELMA lens model to allow for hierarchical judgment tasks. This has been done using the traditional double system lens model (e.g., [135], [136]). A similar extension to demonstrate the scalability of the ELMA model would be useful to investigate hierarchical judgment tasks supported by information automation.

Additional future work could also involve refinement of the definitions (and potentially additions) to the levels of information automation support. The current definitions of the levels could be enhanced to reduce the ambiguity between adjacent levels and to better account for instantiations of automated support that could be argued to fall into one level or another. For example, one could argue that the perceive level of support instantiated in the quality of care study presented in Chapter 5 was actually comprehend support. However, regardless of the nuances in the taxonomy and the difficulty in designing automation that clearly falls into one category of support, ELMA still provides the most systematic framework to date to investigate automation specifically designed to support the cognitive functions involved in human judgment.

One application area that may particularly benefit from the ELMA framework is the design and evaluation of electronic medical records (EMRs). Many judgment tasks are performed by perceiving, comprehending, and integrating cues in order to arrive at a judgment regarding the true state of a patient. A systematic and quantitative approach to informing and investigating the functionality of EMRs that support such tasks is imperative, particularly given the recent widespread adoption of such information automation systems.

References

[1] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, “A model for types and levels of human interaction with automation,” *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[2] D. B. Kaber and M. R. Endsley, “Out-of-the-loop performance problems and the use of intermediate levels of automation for improved control system functioning and safety,” *Process Safety Progress*, vol. 16, no. 3, pp. 126–131, 1997.

[3] D. B. Kaber and M. R. Endsley, “The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task,” *Theoretical Issues in Ergonomics Science*, vol. 5, no. 2, pp. 113–153, 2004.

[4] M. R. Endsley and D. B. Kaber, “Level of automation effects on performance, situation awareness and workload in a dynamic control task,” *Ergonomics*, vol. 42, no. 3, pp. 462–492, 1999.

[5] M. R. Endsley and E. O. Kiris, “The out-of-the-loop performance problem and level of control in automation,” *Human Factors*, vol. 37, no. 2, pp. 381–394, 1995.

[6] R. Parasuraman, “Designing automation for human use: Empirical studies and quantitative models,” *Ergonomics*, vol. 43, no. 7, pp. 931–951, 2000.

[7] E. J. Bass and A. R. Pritchett, “Human-Automated Judge Learning: A methodology for examining human interaction with information analysis automation,” *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 38, no. 4, pp. 759–776, 2008.

[8] W. J. Horrey, C. D. Wickens, R. Strauss, A. Kirlik, and T. R. Stewart, “Supporting situation assessment through attention guidance and diagnostic aiding: The benefits and costs of display enhancement on judgment skill,” in *Adaptive Perspectives on Human-Technology Interaction*, A. Kirlik, Ed. Oxford: Oxford University Press, 2006.

[9] R. Strauss and A. Kirlik, “Situation awareness as judgment II: Experimental demonstration,” *International Journal of Industrial Ergonomics*, vol. 36, no. 5, pp. 475–484, 2006.

[10] A. Kirlik and R. Strauss, “Situation awareness as judgment I: Statistical modeling and quantitative measurement,” *International Journal of Industrial Ergonomics*, vol. 36, no. 5, pp. 463–474, 2006.

[11] A. M. Bisantz and A. R. Pritchett, “Measuring the fit between human judgments and automated alerting algorithms: A study of collision detection,” *Human Factors*, vol. 45, no. 2, pp. 266–280, 2003.

[12] Y. Seong and A. Bisantz, “The impact of cognitive feedback on judgment performance and trust with decision aids,” *International Journal of Industrial Ergonomics*, vol. 38, no. 7–8, pp. 608–625, 2008.

[13] Y. Seong and A. M. Bisantz, “Judgment and trust in conjunction with automated decision aids: A theoretical model and empirical investigation,” in *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomic Society*, 2002, vol. 46, pp. 423–427.

[14] Federal Aviation Administration, “FAA Aerospace Forecast: Fiscal Years 2011-2031,” 2011. [Online]. Available: http://www.faa.gov/about/office_org/headquarters_offices/apl/aviation_forecasts/aerospace_forecasts/2011-2031/media/2011%20Forecast%20Doc.pdf. [Accessed: 25-Jun-2012].

[15] Accreditation Council of Graduate Medical Education, “The ACGME Outcome Project,” *Common Program Requirements*, 2007. [Online]. Available: http://www.acgme.org/outcome/. [Accessed: 14-Feb-2010].

[16] D. A. Davis, P. E. Mazmanian, M. Fordis, R. Van Harrison, K. E. Thorpe, and L. Perrier, “Accuracy of physician self-assessment compared with observed measures of competence: a systematic review,” *Journal of the American Medical Association*, vol. 296, no. 9, pp. 1094–1102, 2006.

[17] J. A. Lyman, J. Schorling, M. Nadkarni, N. May, K. Scully, and J. Voss, “Development of a web-based resident profiling tool to support training in practice-based learning and improvement,” *Journal of General Internal Medicine*, vol. 23, no. 4, pp. 485–488, 2008.

[18] K. L. Mosier and U. M. Fischer, “Judgment and decision making by individuals and teams: Issues, models, and applications,” *Reviews of Human Factors and Ergonomics*, vol. 6, no. 1, pp. 198–256, 2010.

[19] A. M. Bisantz, A. Kirlik, P. Gay, D. A. Phipps, N. Walker, and A. D. Fisk, “Modeling and analysis of a dynamic judgment task using a lens model approach,” *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 30, no. 6, pp. 605–616, 2000.

[20] L. A. Baumgart, E. J. Bass, B. Philips, and K. Kloesel, “Emergency management decision making during severe weather,” *Weather and Forecasting*, vol. 23, no. 6, pp. 1268–1279, 2008.

[21] W. H. Harman, “TCAS: A system for preventing midair collisions,” *The Lincoln Laboratory Journal*, vol. 2, no. 3, pp. 437–457, 1989.

[22] A. A. Montgomery, “Evaluation of computer based clinical decision support system and risk chart for management of hypertension in primary care: Randomised controlled trial,” *British Medical Journal*, vol. 320, no. 7236, pp. 686–690, 2000.

[23] L. A. Baumgart, E. J. Bass, J. A. Lyman, S. Springs, J. Voss, G. F. Hayden, M. A. Hellems, T. R. Hoke, K. A. Schlag, and J. B. Schorling, “Supporting physicians’ practice-based learning and improvement (PBLI) and quality improvement through exploration of population-based medical data,” in *Proceedings of the 54th Annual Meeting of the Human Factors and Ergonomic Society*, 2010, vol. 54, pp. 845–849.

[24] L. T. Kohn, J. M. Corrigan, and M. Donaldson, *To Err Is Human: Building a Safer Health System*. Washington, DC: Institute of Medicine, 1999.

[25] D. Blumenthal, “Launching HITECH,” *The New England Journal of Medicine*, vol. 362, no. 5, pp. 382–385, 2010.

[26] A. Degani, *Taming HAL: Designing interfaces beyond 2001*. New York, NY: Palgrave Macmillan, 2004.

[27] N. Sarter, D. Woods, and C. E. Billings, “Automation surprises,” in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed. New York, NY: Wiley, 1997, pp. 1926–1943.

[28] K. B. Bennett and J. M. Flach, “Graphical displays: implications for divided attention, focused attention, and problem solving,” *Human Factors*, vol. 34, no. 5, pp. 513–533, 1992.

[29] P. M. Fitts, *Human engineering for an effective air-navigation and traffic-control system*. Washington, DC: National Research Council, 1951.

[30] T. B. Sheridan, “Function allocation: algorithm, alchemy or apostasy?,” *International Journal of Human-Computer Studies*, vol. 52, no. 2, pp. 203–216, 2000.

[31] A. Chapanis, “On the allocation of functions between men and machines,” *Occupational Psychology*, vol. 39, pp. 1–11, 1965.

[32] T. B. Sheridan, “Allocating functions rationally between humans and machines,” *Ergonomics in Design: The Quarterly of Human Factors Applications*, vol. 6, no. 3, pp. 20–25, 1998.

[33] A. Dearden, “Allocation of function: scenarios, context and the economics of effort,” *International Journal of Human-Computer Studies*, vol. 52, no. 2, pp. 289–318, 2000.

[34] G. Grote, “KOMPASS: a method for complementary function allocation in automated work systems,” *International Journal of Human-Computer Studies*, vol. 52, no. 2, pp. 267–287, 2000.

[35] R. Fuld, “The fiction of function allocation, revisited,” *International Journal of Human-Computer Studies*, vol. 52, no. 2, pp. 217–233, 2000.

[36] T. B. Sheridan and W. L. Verplank, “Human and computer control of undersea teleoperators,” MIT, Boston, MA, Man Machine Systems Laboratory Report, 1978.

[37] J. Rasmussen, “Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 13, no. 3, pp. 257–266, 1983.

[38] C. Wickens, *An Introduction to Human Factors Engineering*, 2nd ed. Upper Saddle River N.J.: Pearson Prentice Hall, 2004.

[39] M. R. Endsley, “Toward a theory of situation awareness in dynamic systems,” *Human Factors*, vol. 37, no. 1, pp. 32–64, 1995.

[40] R. Cooksey, *Judgment Analysis: Theory, Methods, and Applications*. San Diego, CA: Academic Press, 1996.

[41] E. Brunswik, *Perception and the Representative Design of Psychological Experiments*. University of California Press, 1956.

[42] K. Hammond, *Human Judgment and Decision Making Theories, Methods, and Procedures*. New York, NY: Praeger, 1980.

[43] K. R. Hammond, “Expansion of Egon Brunswik’s psychology, 1955-1995,” in *The Essential Brunswik: Beginnings, Explications, Applications*, K. R. Hammond and T. R. Stewart, Eds. Oxford: Oxford University Press, 2001.

[44] K. R. Hammond, “Probabilistic functionalism and the clinical method,” *Psychological Review*, vol. 62, pp. 255–262, 1955.

[45] J. Smedslund, *Multiple-probability Learning: An Inquiry into the Origins of Perception*. Oslo, Norway: Oslo University Press, 1955.

[46] K. R. Hammond, F. J. Todd, M. Wilkins, and T. O. Mitchell, “Cognitive conflict between persons: Application of the ‘lens model’ paradigm,” *Journal of Experimental Social Psychology*, vol. 2, pp. 343–360, 1966.

[47] K. R. Hammond, M. M. Wilkins, and F. J. Todd, "A research paradigm for the study of interpersonal learning," *Psychological Bulletin*, vol. 65, pp. 221–232, 1966.

[48] K. R. Hammond, T. R. Stewart, B. Brehmer, and D. O. Steinmann, "Social judgment theory," in *Human Judgment and Decision Processes*, M. F. Kaplan and S. Schwartz, Eds. New York, NY: Academic Press, 1975, pp. 272–312.

[49] R. A. Bottenberg and R. E. Christal, "An iterative technique for clustering criteria which retains optimum predictive efficiency," Personnel Research Laboratory, Wright Air Development Division, Lackland Air Force Base, TX, WADD-TN-61-30, ASTIA Document AD-261 615, 1961.

[50] L. R. Goldberg, "Simple models or simple processes? Some research on clinical judgments," *The American Psychologist*, vol. 23, no. 7, pp. 483–496, 1968.

[51] G. E. Campbell, W. L. Buff, and A. E. Bolton, "The diagnostic utility of fuzzy system modeling for application in training systems," in *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomic Society*, 2000, vol. 44, pp. 370–373.

[52] H. J. Einhorn, D. N. Kleinmuntz, and B. Kleinmuntz, "Linear regression and process tracing models of judgment," *Psychological Review*, vol. 86, pp. 465–485, 1979.

[53] R. M. Dawes, "The robust beauty of improper linear models in decision making," *American Psychologist*, vol. 34, pp. 571–582, 1979.

[54] R. W. Cooksey, "Social judgment theory in education: Current and potential applications," in *Human Judgment: The SJT View*, B. Brehmer and C. R. B. Joyce, Eds. Amsterdam: North-Holland, 1988, pp. 273–316.

[55] J. E. Heald, "Social judgment theory: Applications to educational decision making," *Educational Administration Quarterly*, vol. 27, pp. 343–357, 1991.

[56] R. E. Snow, "Brunswikian approaches to research on teaching," *American Educational Research Journal*, vol. 5, pp. 475–489, 1968.

[57] P. Slovic, L. G. Rorer, and P. J. Hoffman, “Analyzing the use of diagnostic drugs,” *Investigative Radiology*, vol. 6, pp. 18–26, 1971.

[58] D. G. Smith and R. S. Wigton, “Research in medical ethics: The role of social judgment theory,” in *Human Judgment: The SJT View*, B. Brehmer and C. R. B. Joyce, Eds. Amsterdam: North-Holland, 1988, pp. 427–442.

[59] R. S. Wigton, “Applications of judgment analysis and cognitive feedback to medicine,” in *Human Judgment: The SJT View*, B. Brehmer and C. R. B. Joyce, Eds. Amsterdam: North-Holland, 1988, pp. 227–246.

[60] R. Libby, *Accounting and Human Information Processing Theory and Applications*. Englewood Cliffs NJ: Prentice-Hall, 1981.

[61] W. S. Waller, “Brunswikian research in accounting and auditing,” in *Human Judgment: The SJT View*, B. Brehmer and C. R. B. Joyce, Eds. Amsterdam: North-Holland, 1988, pp. 247–272.

[62] T. C. Earle and G. Cvetkovich, “Risk judgment, risk communication, and conflict management,” in *Human Judgment: The SJT View*, B. Brehmer and C. R. B. Joyce, Eds. Amsterdam: North-Holland, 1988, pp. 361–400.

[63] L. I. Dalgleish, “Decision making in child abuse cases: Applications of social judgment theory and signal detection theory,” in *Human Judgment: The SJT View*, B. Brehmer and C. R. B. Joyce, Eds. Amsterdam: North-Holland, 1988, pp. 317–360.

[64] T. R. Stewart, W. R. Moninger, R. H. Brady, F. H. Merrem, T. R. Stewart, and J. Grassia, “Analysis of expert judgment in a hail forecasting experiment,” *Weather and Forecasting*, vol. 4, pp. 24–34, 1989.

[65] J. Parkin, *Judging Plans and Projects: Analysis and Public Participation in the Evaluation Process*. Aldershot, England: Avebury, 1993.

[66] D. Brady, “Policy-capturing in the field: The nuclear safeguards problem,” *Organizational Behavior and Human Performance*, vol. 9, pp. 253–266, 1973.

[67] C. J. Hursch, K. R. Hammond, and J. L. Hursch, “Some methodological considerations in multiple-cue probability studies,” *Psychological Review*, vol. 71, no. 1, pp. 42–60, 1964.

[68] L. R. Tucker, “A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch, and Todd,” *Psychological Review*, vol. 71, no. 6, pp. 528–530, 1964.

[69] T. Stewart and C. Lusk, “Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts,” *Journal of Forecasting*, vol. 13, no. 7, pp. 579–599, 1994.

[70] A. D. MacCormick and B. R. Parry, “Judgment analysis of surgeons’ prioritization of patients for elective general surgery,” *Medical Decision Making*, vol. 26, no. 3, pp. 255–264, 2006.

[71] A. H. Murphy, “Skill scores based on the mean square error and their relationships to the correlation coefficient,” *Monthly Weather Review*, vol. 116, no. 12, pp. 2417–2424, 1988.

[72] A. J. Masalonis and R. Parasuraman, “Effects of situation-specific reliability on trust and usage of automated air traffic control decision aids,” in *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomic Society*, 2003, vol. 47, pp. 533–537.

[73] K. A. Klein, “The effects of cognitive feedback on trust and performance while using information analysis automation,” The University of Virginia, Charlottesville, VA, 2005.

[74] B. Lorenz, F. Di Nocera, S. Röttger, and R. Parasuraman, “Automated fault-management in a simulated spaceflight micro-world,” *Aviation, Space, and Environmental Medicine*, vol. 73, no. 9, pp. 886–897, 2002.

[75] C. D. Wickens and X. Xu, “Automation Trust, Reliability and Attention HMI 02 03,” University of Illinois, Aviation Human Factors Division, Savoy, IL, AHDF Technical Report AHFD-02- 14/MAAD-02-2, 2002.

[76] C. D. Wickens, H. Li, A. Santamaria, A. Sebok, and N. B. Sarter, “Stages and levels of automation: An integrated meta-analysis,” in *Proceedings of the 54th Annual Meeting of the Human Factors and Ergonomic Society*, 2010, vol. 54, pp. 389–393.

[77] W. M. Crocoll and B. G. Coury, “Status or recommendation: Selecting the type of information for decision aiding,” in *Proceedings of the 34th Annual Meeting of the Human Factors and Ergonomic Society*, 1990, vol. 34, pp. 1524–1528.

[78] E. Rovira, K. McGarry, and R. Parasuraman, “Effects of imperfect automation on decision making in a simulated command and control task,” *Human Factors*, vol. 49, no. 1, p. 76, 2007.

[79] N. B. Sarter and B. Schroeder, “Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing,” *Human Factors*, vol. 43, no. 4, pp. 573–583, 2001.

[80] M. C. Wright and D. B. Kaber, “Effects of automation of information-processing functions on teamwork,” *Human Factors*, vol. 47, no. 1, pp. 50–66, 2005.

[81] F. A. Drews and D. R. Westenskow, “The right picture is worth a thousand numbers: Data displays in anesthesia,” *Human Factors*, vol. 48, no. 1, pp. 59–71, 2006.

[82] K. J. Vicente, “Ecological interface design: Progress and challenges,” *Human Factors*, vol. 44, no. 1, pp. 62–78, 2002.

[83] E. L. Wiener, “Cockpit automation,” in *Human Factors in Aviation*, E. L. Wiener and D. C. Nagel, Eds. New York: Academy Press, 1988, pp. 433–461.

[84] A. Kirlik, N. Walker, A. D. Fisk, and K. Nagel, “Supporting perception in the service of dynamic decision making,” *Human Factors*, vol. 38, no. 2, pp. 288–299, 1996.

[85] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick, “Automation bias: decision making and performance in high-tech cockpits,” *The International Journal of Aviation Psychology*, vol. 8, no. 1, pp. 47–63, 1997.

[86] M. Yeh, C. D. Wickens, and F. J. Seagull, “Target cuing in visual search: the effects of conformality and display location on the allocation of visual attention,” *Human Factors*, vol. 41, no. 4, pp. 524–542, 1999.

[87] U. Metzger and R. Parasuraman, “Conflict detection aids for air traffic controllers in free flight: Effects of reliable and failure modes on performance and eye movements,” in *Proceedings of the 11th International Symposium on Aviation Psychology*, Columbus, OH, 2001.

[88] A. Kirlik, “Requirements for psychological models to support design: Towards ecological task analysis,” in *Global Perspectives on the Ecology of Human–Machine Systems*, vol. 1, J. M. Flach, P. A. Hancock, J. K. Caird, and K. J. Vicente, Eds. Hillsdale, NJ: Lawrence Erlbaum, 1995, pp. 68–120.

[89] K. J. Vicente and J. Rasmussen, “The ecology of human-machine systems II: Mediating ‘direct perception’ in complex work domains,” *Ecological Psychology*, vol. 2, no. 3, pp. 207–249, 1990.

[90] D. D. Woods, “The cognitive engineering of problem representation,” in *Human–Computer Interaction in Complex Systems*, G. R. S. Weir and J. L. Alty, Eds. New York: Academic Press, 1991, pp. 169–188.

[91] C. Layton, P. J. Smith, and E. McCoy, “Design of a cooperative problem-solving system for en-route flight planning: An empirical evaluation,” *Human Factors*, vol. 36, no. 1, pp. 94–119.

[92] S. D. Davis and A. R. Pritchett, “Alerting system assertiveness, knowledge, and over-reliance,” *Journal of Information Technology Impact*, vol. 1, no. 3, pp. 119–143, 1999.

[93] T. R. Stewart, K. F. Heideman, W. R. Moninger, and P. Reagan-Cirincione, “Effects of improved information on the components of skill in weather forecasting,” *Organizational Behavior and Human Decision Processes*, vol. 53, no. 2, pp. 107–134, 1992.

[94] R. Parasuraman and V. Riley, “Humans and automation: Use, misuse, disuse, abuse,” *Human Factors*, vol. 39, no. 2, 1997.

[95] M. Lewis, “Designing for human-agent interaction,” *Artificial Intelligence Magazine*, vol. 19, no. 1, pp. 67–78, 1998.

[96] A. R. Pritchett and B. Vandor, “Designing situation displays to promote conformance to automatic alerts,” in *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomic Society*, 2001, vol. 45, pp. 311 – 315.

[97] F. J. Seagull and P. M. Sanderson, “Anesthesia alarms in context: An observational study,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 43, no. 1, pp. 66–78, Jan. 2001.

[98] C. D. Wickens, A. Mavor, R. Parasuraman, and J. McGee, *The future of air traffic control: Human operators and automation*. Washington, D.C.: National Academies Press, 1998.

[99] J. K. Kuchar and L. C. Yang, “A review of conflict detection and resolution modeling methods,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 1, no. 4, pp. 179–189, 2000.

[100] N. B. Sarter and D. D. Woods, “Pilot interaction with cockpit automation: Operational experiences with the flight management system,” *The International Journal of Aviation Psychology*, vol. 2, no. 4, pp. 303–321, 1992.

[101] N. B. Sarter and D. D. Woods, “Pilot interaction with cockpit automation II: An experimental study of pilots’ model and awareness of the flight management system,” *The International Journal of Aviation Psychology*, vol. 4, no. 1, pp. 1–28, 1994.

[102] B. Brehmer, “Note on clinical judgment and the formal characteristics of clinical tasks,” *Psychological Bulletin*, vol. 83, no. 5, pp. 778–782, 1976.

[103] E. J. Bass, “Human-Automated Judgment Learning: A research paradigm based on interpersonal learning to investigate human interaction with automated judgments of hazards,” Unpublished doctoral dissertation., Georgia Institute of Technology, Atlanta, GA, 2002.

[104] W. Balzer, “Task information, cognitive information, or functional validity information: Which components of cognitive feedback affect performance?,” *Organizational Behavior and Human Decision Processes*, vol. 53, no. 1, pp. 35–54, 1992.

[105] W. K. Balzer, M. E. Doherty, and R. O’Connor, “Effects of cognitive feedback on performance,” *Psychological Bulletin*, vol. 106, no. 3, pp. 410–433, 1989.

[106] W. K. Balzer, L. M. Sulsky, L. B. Hammer, and K. E. Sumner, “Task information, cognitive information, or functional validity information: Which components of cognitive feedback affect performance?,” *Organizational Behavior and Human Decision Processes*, vol. 53, no. 1, pp. 35–54, 1992.

[107] G. Gattie and A. Bisantz, “The effects of integrated cognitive feedback components and task conditions on training in a dental diagnosis task,” *International Journal of Industrial Ergonomics*, vol. 36, no. 5, pp. 485–497, 2006.

[108] E. R. Tufte, *Visual Explanations*. Cheshire, CT: Graphic Press, 1997.

[109] E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd ed. Cheshire, CT: Graphic Press, 2001.

[110] J. Nielsen and R. L. Mack, *Usability Inspection Methods*. Ann Arbor, MI: Wiley, 1994.

[111] A. R. Pritchett, “Display effects on shaping apparent strategy: A case study in collision detection and avoidance,” *The International Journal of Aviation Psychology*, vol. 10, no. 1, pp. 59–83, 2000.

[112] D. T. Bauer, S. Guerlain, and P. J. Brown, “The design and evaluation of a graphical display for laboratory data,” *Journal of the American Medical Informatics Association*, vol. 17, no. 4, pp. 416–424, 2010.

[113] J. Cushing, L. D. Janssen, S. Allen, and S. Minocha, “Overview+detail in a tomahawk mission-to-platform assignment tool: Applying information visualization in support of an asset allocation planning task,” *Information Visualization*, vol. 5, no. 1, pp. 1–14, 2006.

[114] G. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building*. New York, NY: Wiley, 1978.

[115] G. Gattie and A. Bisantz, “The effects of integrated cognitive feedback components and task conditions on training in a dental diagnosis task,” *International Journal of Industrial Ergonomics*, vol. 36, no. 5, pp. 485–497, 2006.

[116] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, “The role of trust in automation reliance,” *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[117] J. Lee and N. Moray, “Trust, self-confidence, and operators’ adaptation to automation,” *International Journal of Human-Computer Studies*, vol. 40, no. 1, pp. 153–184, 1994.

[118] P. de Vries, C. Midden, and D. Bouwhuis, “The effects of errors on system trust, self-confidence, and the allocation of control in route planning,” *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 719–735, 2003.

[119] Centers for Medicare and Medicaid Services, “Physician Quality Reporting System formerly known as the Physician Quality Reporting Initiative,” 20-Apr-2011. [Online]. Available: http://www.cms.gov/PQRS//. [Accessed: 08-Jun-2011].

[120] American Board of Medical Specialties, “ABMS Maintenance of Certification,” 2006. [Online]. Available: http://www.abms.org/Maintenance_of_Certification/ABMS_MOC.aspx. [Accessed: 14-Feb-2010].

[121] S. I. Wasserman, H. R. Kimball, and F. D. Duffy, “Recertification in internal medicine: A program of continuous professional development. Task Force on Recertification,” *Annals of Internal Medicine*, vol. 133, no. 3, pp. 202–208, 2000.

[122] B. J. Wu, P. A. Dietz, J. Bordley, and D. C. Borgstrom, “A novel, web-based application for assessing and enhancing practice-based learning in surgery residency,” *Journal of Surgical Education*, vol. 66, no. 1, pp. 3–7, 2009.

[123] G. Ogrinc, L. A. Headrick, L. J. Morrison, and T. Foster, “Teaching and assessing resident competence in practice-based learning and improvement,” *Journal of General Internal Medicine*, vol. 19, no. 5 pt 2, pp. 496–500, 2004.

[124] J. J. Mohr, G. D. Randolph, M. M. Laughon, and E. Schaff, “Integrating improvement competencies into residency education: a pilot project from a pediatric continuity clinic,” *Ambulatory Pediatrics*, vol. 3, no. 3, pp. 131–136, 2003.

[125] M. Srinivasan, K. E. Hauer, C. Der-Martirosian, M. Wilkes, and N. Gesundheit, “Does feedback matter? Practice-based learning for medical students after a multi-institutional clinical performance examination,” *Medical Education*, vol. 41, no. 9, pp. 857–865, 2007.

[126] J. L. Paukert, H. S. Chumley-Jones, and J. H. Littlefield, “Do peer chart audits improve residents’ performance in providing preventive care?,” *Academic Medicine*, vol. 78, no. 10, pp. S39–S41, 2003.

[127] K. W. Eva and G. Regehr, “Exploring the divergence between self-assessment and self-monitoring,” *Advances in Health Sciences Education: Theory and Practice*, 2010.

[128] L. Baumgart, E. J. Bass, J. A. Lyman, and J. Voss, “Supporting the investigation of clinical practice through interactive population-based reporting,” presented at the 2012 Symposium on Human Factors and Ergonomics in Health Care: Bridging the Gap, Baltimore, MD, 2012.

[129] D. Lloyd-Jones, R. Adams, M. Carnethon, G. De Simone, T. B. Ferguson, K. Flegal, E. Ford, K. Furie, A. Go, K. Greenlund, N. Haase, S. Hailpern, M. Ho, V. Howard, B. Kissela, S. Kittner, D. Lackland, L. Lisabeth, A. Marelli, M. McDermott, J. Meigs, D. Mozaffarian, G. Nichol, C. O’Donnell, V. Roger, W. Rosamond, R. Sacco, P. Sorlie, R. Stafford, J. Steinberger, T. Thom, S. Wasserthiel-Smoller, N. Wong, J. Wylie-Rosett, and Y. Hong, “Heart disease and stroke statistics-2009 update: A report

from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee," *Circulation*, vol. 119, no. 3, pp. e21–181, 2009.

[130] A. V. Chobanian, G. L. Bakris, H. R. Black, W. C. Cushman, L. A. Green, J. L. Izzo Jr, D. W. Jones, B. J. Materson, S. Oparil, J. T. Wright Jr, and E. J. Roccella, "The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure: the JNC 7 report," *The Journal of the American Medical Association*, vol. 289, no. 19, pp. 2560–2572, May 2003.

[131] P. Winston, *Artificial intelligence*, 3rd ed. Reading, MA: Addison-Wesley Pub. Co., 1992.

[132] L. A. Zadeh, "Fuzzy logic," *Computer*, vol. 21, no. 4, pp. 83–93, 1988.

[133] P. M. Jones, R. W. Chu, and C. M. Mitchell, "A methodology for human-machine systems research: Knowledge engineering, modeling, and simulation," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 25, no. 7, pp. 1025–1038, 1995.

[134] G. A. Klein, "Applied decision making," in *Handbook of Perception and Cognition: Human Performance and Ergonomics*, P. A. Hancock, Ed. New York, NY: Academic Press, 1999.

[135] R. W. Cooksey and P. Freebody, "Generalized multivariate lens model analysis for complex human inference tasks," *Organizational Behavior and Human Decision Processes*, vol. 35, no. 1, pp. 46–72, 1985.

[136] R. W. Cooksey and P. Freebody, "Cue subset contributions in the hierarchical multivariate lens model: Judgments of children's reading achievement," *Organizational Behavior and Human Decision Processes*, vol. 39, no. 1, pp. 115–132, 1987.