Measurable Decision Making with GSR and Pupillary Analysis for Intelligent User Interface

JIANLONG ZHOU, JINJUN SUN, FANG CHEN, YANG WANG, RONNIE TAIB, AHMAD KHAWAJI, and ZHIDONG LI, National ICT Australia

This article presents a framework of adaptive, measurable decision making for Multiple Attribute Decision Making (MADM) by varying decision factors in their types, numbers, and values. Under this framework, decision making is measured using physiological sensors such as Galvanic Skin Response (GSR) and eye-tracking while users are subjected to varying decision quality and difficulty levels. Following this quantifiable decision making, users are allowed to refine several decision factors in order to make decisions of high quality and with low difficulty levels. A case study of driving route selection is used to set up an experiment to test our hypotheses. In this study, GSR features exhibit the best performance in indexing decision quality. These results can be used to guide the design of intelligent user interfaces for decision-related applications in HCI that can adapt to user behavior and decision-making performance.

Categories and Subject Descriptors: H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—*Evaluation/methodology*

General Terms: Design, Human Factors, Experimentation

Additional Key Words and Phrases: Decision making, GSR, eye-tracking, machine learning

ACM Reference Format:
Jianlong Zhou, Jinjun Sun, Fang Chen, Yang Wang, Ronnie Taib, Ahmad Khawaji, and Zhidong Li. 2015. Measurable decision making with GSR and pupillary analysis for intelligent user interface. ACM Trans. Comput.-Hum. Interact. 21, 6, Article 33 (January 2015), 23 pages.
DOI: http://dx.doi.org/10.1145/2687924

1. INTRODUCTION

In recent years, decision making has become an important topic in various areas of Human-Computer Interaction (HCI) research [Smith et al. 2009]. Additionally, non-verbal information, such as physiological information, is increasingly parsed and interpreted by computers to interactively construct and refine models of human cognitive and affective states [Stickel et al. 2009; Wang et al. 2013]. Such user models, together with Machine Learning (ML) techniques, can then be used in an adaptive fashion to enhance HCIs and make interfaces appear intelligent [Duric et al. 2002; Holzinger 2013]. Therefore, the use of physiological measurements in decision making promises to provide a rich and enduring approach to building intelligent HCI systems that adapt to users' behavior and their decision making performance. Imagine a computer interface that could predict and diagnose whether a decision made by a user corresponded to a

This work is partly supported by the Asian Office of Aerospace Research & Development (AOARD) under grant No. FA2386-14-1-0007 AOARD 134144, and FA2386-14-1-0022 AOARD 134131.
Authors' address: J. Zhou, J. Sun, F. Chen, Y. Wang, R. Taib, A. Khawaji, and Z. Li, National ICT Australia (NICTA), Level 5, 13 Garden St., Eveleigh, NSW 2015, Australia; emails: {jianlong.zhou, fang.chen, yang.wang, ronnie.taib, ahmad.khawaji, zhidong.li}@nicta.com.au; jsunster@gmail.com.

DOI: http://dx.doi.org/10.1145/2687924

low difficulty level and resulted in a high-quality decision by simply collecting a variety of physiological information (e.g., pupillary responses, skin responses) from the user. Further imagine that the interface could adaptively vary decision factors during decision tasks to improve decision quality—then imagine the resulting HCI improvements possible from using these predictions and diagnoses.

Decisions are made on various aspects of life every day. Every decision-making process produces a final choice of action or opinion among several alternative scenarios [Juliusson et al. 2005]. A decision-making process can be divided into two parts: declaring a decision and working a decision. During the declaration of a decision, one first needs to frame the problem of what is to be decided upon. Then one needs to define who is to be involved in the decision, followed by declaring what approaches are to be used to make the decision. While working out a decision, one needs to formulate a complete set of alternatives. After that, values are defined to allow tradeoffs to be made between alternatives. Finally, information that describes the value of each alternative is analyzed to make a decision. This could be an iterative process.

Different alternatives in decision making are often characterized by multiple attributes. We also refer to these attributes as *decision factors* in this article. Various attributes usually have different units, such as the decision factors used in travel route decision making: length, travel time, speed, and the like. Multiple Attribute Decision Making (MADM) involves "making preference decisions over the available alternatives that are characterized by multiple, usually conflicting, attributes" [Azar 2000; Kim et al. 2012]. MADM is intrinsically difficult because multiple attributes can conflict with each other. This makes the selection task even harder since it is very difficult to get an applicable model to calculate the value of each alternative [Kim et al. 2012]. Therefore, MADM often involves high cognitive load [Bettman et al. 1990]. However, little work in MADM is done in evaluating how variations in attributes, such as types and values, affect decision making.

Because MADM often involves overwhelming information and laborious cognitive processes, it would be interesting to open a window into a person's thinking while he or she investigates decision factor variations in MADM. According to the "eye-mind hypothesis," eye tracking results can reveal the underlying cognitive processes of a human user [Chen et al. 2011]. Thus, the eye is literally the window to the mind. Much work has been done on using eye tracking to understand the human decision-making process [Preuschoff et al. 2011; Fiedler and Glockner 2012]. However, little work exists that studies pupillary responses in decision making involving decision factor variations. Furthermore, it was found that Galvanic Skin Response (GSR), which corresponds to the electrical conductance of the skin, as a low-cost and robust physiological signal, accurately reflects the process of decision making—particularly emotional sanctioning of an active go-ahead [Dawson et al. 2011; Boucsein 2012]. GSR is often used as an indicator of affective processes and emotional arousal. Therefore, in addition to the analysis of pupillary responses, we also analyze GSR during decision making in order to investigate how physiological signals are used to index such decision making under varying decision factors.

To fill gaps in the salient research, this article suggests that decision making can be measured in real time in order to let users perceive the quality of their decisions and the level of difficulty involved in making these decisions. Armed with such a performance indicator of their decision making, users can refine the decision factors utilized to arrive at high-quality decisions with low difficulty levels. Following this concept, a framework of adaptive measurable decision making is proposed by varying decision factors in their types, numbers, and values. Under this framework, decision making is measured using physiological sensors such as GSR and an eye tracker. The framework presents a novel intelligent interface in which human and computer can

augment each other's capabilities. The framework analyzes the physiological signals of users with computational algorithms. These in turn feed into processes that adapt decision factors in the user interface to enhance user performance in decision making. A case study of driving route selection is used to set up an experiment to test our hypotheses. The study results can be used to guide the design of a user interface for decision-related applications in HCI. The proposed framework is viewed as a necessary step in developing intelligent HCI systems where human physiological information is modeled and used to adapt both interface and decision making. In summary, the overall objectives of this study include:

—Propose a framework of adaptive measurable decision making;
—Demonstrate that decision making can be measured quantitatively to let users perceive both the quality of their decisions and the level of difficulty of the decision process based on physiological signals;
—Explore decision making quantitatively by varying decision factors' type, number, and values.

In the following sections, Section 2 introduces related work on decision making and physiological signals in decision making. Section 3 poses the hypotheses of the study. Section 4 presents a framework of adaptive measurable decision making and introduces a visualization approach used to present multiple decision factors in decision making. A case study is also introduced in this section. Section 5 sets up the experiment based on the case study. We analyze subjective ratings of the experiment in Section 6. GSR and pupillary responses are analyzed in Sections 7 and 8, respectively. We discuss the significant findings of this study in Section 9 before concluding the paper in Section 10.

2. RELATED WORK

2.1. Perspectives on Decision Making

Extensive research has been done in MADM [Lertprapai 2013; Shin et al. 2013]. A typical MADM problem involves a number of alternatives to be assessed and a number of criteria to assess the alternatives. Each alternative has a value for each attribute, and, based on these values, the alternatives can be assessed and ranked [Lertprapai 2013; Shin et al. 2013]. Various models are developed to assess multiple attributes, such as the Borda-Kendall model [Kendall 1962] for ordinal preference measurements, the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) [Hwang and Yoon 1981], and the Simple Additive Weighting (SAW) approach [MacCrimon 1968]. In these approaches, multiple attributes can be expressed in a matrix format, where rows represent alternatives and columns represent the attributes considered [Hwang and Yoon 1981]. It was found that a greater number of decision options may hinder a participant's motivation to make a decision [Sethi-Iyengar et al. 2004; McWilliams et al. 2011]. Furthermore, in order to deal with problems having both quantitative and qualitative attributes in MADM, the *evidential reasoning approach* [Yang and Sen 1994; Yang and Xu 2013] has been developed to analyze decisions with multiple attributes. This approach is based on decision theory, particularly *utility theory* [Neumann 1944; Gutnik et al. 2006] and statistical analysis.

Many factors influence decision making. Classical models of decision making usually focus on cognitive, situational, and sociocultural variables in accounting for human performance. Significant factors include past experiences [Juliusson et al. 2005]; a variety of cognitive biases [Stanovich and West 2008]; individual differences [Bruine de Bruin et al. 2007], including age and socioeconomic status; and a belief in personal relevance [Acevedo and Krueger 2004]. Furthermore, an emotional component also accounts for human decision making as a separate factor [Gutnik et al. 2006].

Despite these extensive studies in MADM, little attention is paid to how the types and values of decision attributes affect decision making. Little research has investigated the setup of connections between physiological indicators and MADM with variations in different aspects of decision factors (e.g. types, numbers, and values of decision factors). In our work, physiological measurements are recorded during decision making under various conditions to set up relations between decision making and physiological indicators. We also use a similar method with the evidential reasoning approach to analyze decision performance based on utility and additional physiological indicators.

2.2. Eye Tracking in Decision Making

Eye movements can help researchers evaluate the decision-making process by providing additional insight into the cognitive mechanisms (even at the neurological level) that produce them. Fiedler and Glockner [2012] utilize eye tracking to analyze the dynamics of decision making in risk conditions featuring two gambles. Their work shows that attention to the outcome of a gamble increases with its probability and its value and that attention shifts toward the subsequently favored gamble after about two-thirds of the decision process, indicating a gaze-cascade effect. Pupil dilation, which reflects both cognitive effort and arousal, increases during the decision-making process. A recent investigation [Franco-Watkins and Johnson 2011] also shows that pupil dilation increases over the course of decision making and is influenced by presentation format (basic eye tracking vs. decision moving-window). Another eye tracking study using a card gambling task shows that pupil dilation reflected surprise but not expected reward in decision making [Preuschoff et al. 2011]. Pupillary analysis is also widely used to index cognitive workload [Xu et al. 2011; Wang et al. 2013]. It has been found that pupillary features such as pupil diameter can be used to indicate levels of cognitive workload. Therefore, pupillary response can be used as an objective indicator to measure users' physiological states during decision making.

2.3. GSR in Decision Making

GSR, also called Skin Conductance Response (SCR), is a robust physiological signal that can be measured relatively cheaply, easily, and unobtrusively. It refers to how well the skin conducts electricity when an external direct current of constant voltage is applied, and it is measured in microsiemens [Figner and Murphy 2011]. It yields a continuous measure that is related to activity in the sympathetic branch of the autonomic nervous system. Changes in skin conductance are related to the activity of eccrine sweat glands, which are innervated by sympathetic nerves. These changes reflect the secretion of sweat from these glands. Because sweat is an electrolyte solution, the more the skin's sweat ducts and pores are filled with sweat, the more conductive the skin becomes. It has been found that skin conductance is closely related to various neural and psychological activities in humans [Figner and Murphy 2011]. It is well established that skin conductance covaries with the arousal dimension of affect, thus indexing its intensity.

Rotenberg and Vedenyapin [1985] demonstrates that GSR is particularly pronounced when a decision is being made under conditions where action is delayed. The findings show that the GSR is associated with the process of decision making because it increases when a decision not to act is being made, as well as when a decision to act is being made. The studies by the Iowa group were pioneering in their use of GSR to investigate questions related to decision making. Research using the Iowa Gambling Task (IGT) [Bechara et al. 1994, 1999] demonstrates that GSR can be used as a process indicator of affective processes before, during, and after making decisions. It was found that GSR can distinguish between "good" and "bad" decisions from studies with patients having no diagnosis of brain damage [Boucsein 2012]. Payne [2008] utilized

GSR as a marker of intuitive decision making in nursing. Botvinick and Rosen [2009] observed an anticipatory GSR prior to selection actions in decision making resulting in a high level of cognitive demand. Dawson et al. [2011] demonstrated that GSR can reflect conscious and unconscious emotional processes or serve as a covert physiological marker that guides future decision making. Therefore, GSR can serve as an objective, nonverbal, nonvoluntary indicator and a physiological measure that is relatively free from demand characteristics and reporting biases in decision making.

3. HYPOTHESES

In this study, the following hypotheses are posed:

—**H1:** Decision making with/without certain decision factors will result in differences in the ease of decision making, which will also result in differences of physiological measurements.
—**H2:** The number of certain decision factors will affect the ease of decision making, which will be shown in differences of physiological measurements;
—**H3:** The different values of certain decision factors will result in differences in the ease of decision making, which will be seen in differences of physiological measurements;
—**H4:** Different types of decision factors will result in differences in the ease of decision making, which will also result in differences of physiological measurements.

4. METHOD

This section presents a framework of adaptive measurable decision making. Furthermore, because the presentation of multiple decision factors significantly affects performance of decision making [Kim et al. 2012], this section also proposes a visualization method to present multiple decision factors effectively.

4.1. Framework of Adaptive Measurable Decision Making

We present a framework of adaptive measurable decision making in Figure 1. In this framework, when an experimental task with certain decision factors is presented to users for decision making, task difficulty is measured simultaneously with subjective ratings and physiological measurements (e.g., GSR, eye tracking). After the user makes decisions, the decision performance is evaluated with the user's choice and physiological measurements. The measured information is then analyzed, and classifiers for decision quality and task difficulty are derived.

When a new task is presented, users' workload during decision making is recorded using real-time physiological measurements. These measurements are sent to classifiers to determine the difficulty levels and decision quality learned in the experimental task stage. The task difficulty level and decision quality ratings from classifiers are exported to users. If users are satisfied with their decision performance and decision itself, then the decision-making process is finished. Otherwise, decision factors are refined (e.g., increase/decrease number of certain decision factors) based on the analysis results at the experimental task stage to initiate a new decision-making session. For example, if the decision difficulty level derived from measurements is low, more of certain decision factors may be included in order to produce higher quality decisions. This process is iteratively performed until decisions meet the performance and difficulty requirements of users. In this process, questionnaires are not needed to evaluate decision difficulty levels and quality. Such a framework allows users to refine their decisions adaptively and to interact with the HCI system more efficiently. This article focuses on the investigation of the experimental task stage in Figure 1 (upper part). The new

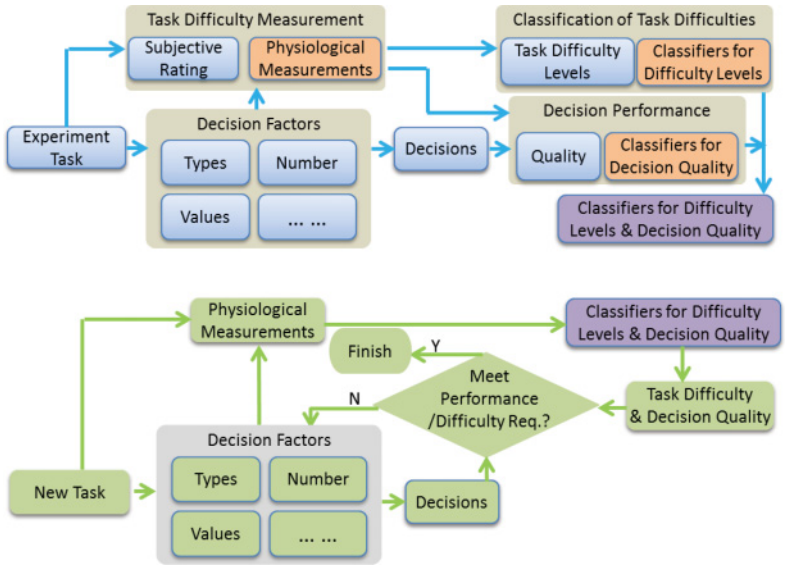


Fig. 1. Framework of adaptive measurable decision making.

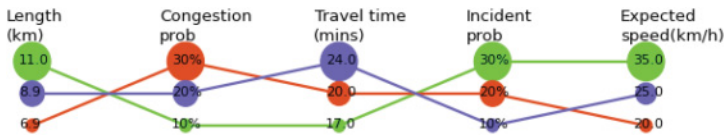


Fig. 2. Parallel SimulSort (PSS) is used to present various decision factors.

task stage in Figure 1 (lower part) will be directly solved after classifiers for difficulty levels and decision quality are received from the experimental task stage.

4.2. Visualization of Decision Factors

To alleviate difficulties in MADM, various visualization techniques have been used to help the decision-making process by making decision factors more interpretable. *Parallel coordinates* is one classical approach to visualizing multiattribute data points. One of advantages of this technique is that it can provide an overview of data trends, which may help in MADM when each attribute is represented as one axis in a parallel coordinate system. One obvious disadvantage of this technique is that it lacks a tabular view for presenting the value details of each coordinates. SimulSort [Kim et al. 2012] organizes different decision factors of all alternatives in a tabular form and sorts all columns simultaneously. However, users still need to undertake laborious interactions in SimulSort to highlight different alternatives for comparison. SimulSort also cannot provide an overall overview of data trends.

In this article, we propose a visualization approach for presenting multiple decision factors by combining the advantages of both parallel coordinates and SimulSort, a method called Parallel SimulSort (PSS) and shown in Figure 2. Similarly to SimulSort, PSS implicitly organizes various decision factors in a tabular form. Various routes are organized as rows in the table; different decision factors are organized as columns and sorted in descending order. Various decision factors belonging to an alternative are

encoded with colors. All decision factors belonging to different alternatives are encoded with different colors, instead of highlighting decision factor alternatives one at a time, as in SimulSort. This color encoding approach allows decision factors belonging to each alternative to form tabular cell-based polylines that provide an overview of alternative data trends, just as parallel coordinates do. The benefit of providing an overview of alternatives and their associated attribute details can improve information browsing efficiency.

4.3. Case Study

This research uses driving travel route decision making as a case study. Travel route decision making is increasingly significant in modern life, both for individuals and society, in ways that extend far beyond driver inconvenience. For example, because of congestion on roads, the cost of wasted fuel and lost productivity reached $100 billion in 2010—more than $750 for every U.S. traveler; the amount of wasted time totaled 4.8 billion hours—34 hours for every traveler in the United States [Xerox 2013].

Various factors may affect travel route decision making [Knorring 2003]. A brief list of these decision factors includes such things as availability of alternate routes, length of alternate routes, perceived speeds on alternate routes, anticipated congestion, hazards avoided (incident), travel time of each route, weather on alternate routes, and scenery encountered. In addition, socioeconomic and other factors, such as income, education, or journey purpose, also affect route decision making [Abdel-Aty et al. 1994]. In this study, route length, travel time, congestion probability, incident probability, and expected speed on alternate routes are chosen to investigate how these factors affect the decision-making process. These decision factors are divided into two groups. (1) *Common decision factors* are common to all decision tasks and are decided by the route itself. Route length and travel time were defined as common decision factors in this study. (2) *Variable decision factors* are different depending on tasks. Variable decision factors in this study were derived from driving history data on the route using ML models, and we refer to them as *ML-based decision factors*. Of course, data from other sources such as reports, gauges, and situation assessment could also be used as values for variable decision factors. In this study, congestion probability, incident probability, and expected speed were defined as variable decision factors. Congestion probability is the probability of anticipated congestion on this route. Its value is in the range of [0.0, 1.0], and the lower, the better. Incident probability is the probability of a possible incident on the route. If the incident rate is close to 0, the route is safer. Despite expected speed possibly being computed as route length over travel time, the expected speed in this study is more related with acceleration and deceleration operations that consume more gasoline. For expected speed, the higher, the better because higher speed may require few acceleration and deceleration operations and thus consume less gasoline.

5. EXPERIMENT SETUP

This section sets up an experiment to test our hypotheses with the case study of travel route decision making.

5.1. Experiment Data

Driving travel route data (e.g., route length, travel time, travel route maps) in various world cities were collected from Google Maps. World cities were from Europe and America, and participants in Australia confirmed that they were not familiar with those city routes and did not have driving experiences in those cities, which allowed us to avoid bias. Simulated data were used in this study as variable decision factors (e.g., congestion probability, incident probability, and expected speed). Travel routes were

Table I. Task Setup in the Experiment

Task#	Routes	Common Factors	Variable Factors
1	3	Length, Travel time	Not available
2	3	Length, Travel time	Congestion: 0.1, 0.2, 0.3
3	3	Length, Travel time	Congestion: 0.1, 0.2, 0.3 Incident: 0.1, 0.2, 0.3
4	3	Length, Travel time	Congestion: 0.1, 0.2, 0.3 Incident: 0.1, 0.2, 0.3 Expected speed: 50, 60, 70
5	3	Length, Travel time	Incident: 0.1, 0.2, 0.3
6	3	Length, Travel time	Congestion: 0.2, 0.2, 0.3
7	3	Length, Travel time	Congestion: 0.3, 0.3, 0.3

from four world cities. Three alternate routes from location A to location B in each city, with various attribute values, were used to allow participants to make a decision. For end users, ML models are black boxes, and users only get ML results in the form of various numbers from ML analyses. Therefore, we can assume that the simulated data functioned in the same role as real ML results for participants in the experiment. As a result, the use of simulated data did not affect the effectiveness of the use of ML results as decision attributes in this work. The use of ML results as conventional decision factors also did not affect the evaluation of decision making in this study.

Participates were told that route length and travel time were common decision factors that did not depend on individual participants. Participants were also told that congestion probability, incident probability, and expected speed were decision factors that were learned from drivers' driving history data on those routes using ML models so that participants would be aware that these decision factors were closely related to routes' actual states.

5.2. Task Design

In this study, each participant acted as a car driver and was supposed to go from location A to location B. The objective basically required each participant to select a route from A to B under the condition of various decision factors. Participants needed to select one target route with the highest score according to all given decision factors, not according to a single decision factor only. There were three routes available from A to B.

As mentioned, five decision factors (two common decision factors of length and travel time; three variable factors of congestion probability, incident probability, and expected speed) were used. Controlled values were applied to variable decision factors as shown in Table I. To test our hypotheses, seven tasks were designed as follows: (1) Three tasks with/without certain decision factor were determined to test H1, (2) three tasks with an increased number of decision factors were determined to test H2, (3) three tasks with different values of a certain decision factor were determined to test H3, and (4) two tasks with different decision factors were determined to test H4. All tasks were combined, and repeated tasks were removed. We finally obtained seven different tasks, as illustrated in Table I. Among these seven tasks, Task 1, Task 2, and Task 3 were used to test H1; Task 2, Task 3, and Task 4 were used to test H2; Task 2, Task 6, and Task 7 were used to test H3; and Task 2 and Task 5 were used to test H4. These seven tasks were performed in each round, and four rounds were performed by each participant (the first round was used as the training round and was not included in the final data

analysis). Therefore, a total of 28 tasks were performed by each participant. The task configurations are shown in Table I ("Routes" refers to number of routes; numbers in the "Variable Factors" column are congestion probabilities, incident probabilities, and/or speeds on each route).

At the beginning of each decision-making task session, a blank screen with a background color identical to that presented in the task session was displayed for 6 seconds to allow the participant to rest and "reset" his or her cognitive load state. An X was then displayed at the center of the blank screen for 3 seconds to further "release" participants' cognitive load state [Wang et al. 2013; Luo and Taib 2013]. Therefore, a total of 9 seconds was allowed to "reset" participant's cognitive state before a map with three alternative routes was displayed. After 3 seconds, the visualization of various decision factors for each route (see a visualization example in Figure 2) was displayed at the bottom of the map until the participant made a decision.

5.3. Participants and Apparatus

Fourteen participants were recruited for the experiment, ranging in age from 20 to 40+. Participants were researchers (including research students) in software, network, and machine learning, and administrative staff, with various levels of education, including bachelor degrees and Ph.Ds. All participants have driving experience.

A GSR device from ProComp Infiniti of Thought Technology Ltd. was used to collect participant skin conductance responses. An eye tracker device from SensoMotoric Instruments GmbH (SMI) was used to collect participant pupillary responses. GSR sensors were attached to participants' left-hand fingers. All participants were right-handed. Travel route maps and decision factors were presented on a 21-inch Dell monitor with a screen resolution of 1024×768 pixels. Figure 3 presents the setup of the experiment and a screenshot of a task performed in the study.

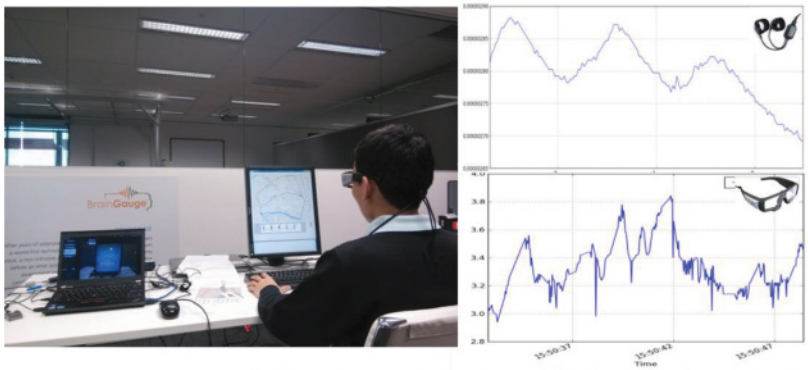
5.4. Data Collection

After each decision-making task session, participants were asked to rate the difficulty level of the task using a 9-point Likert scale (1 = least difficult, and 9 = most difficult). At the end of each round, participants were also asked to rate how important certain decision factors were in making users more confident on their final decisions. Participants were also asked to rate how important different values (e.g., high, average, low) of a particular variable decision factor were in users' deciding to favor a particular route and whether a greater number of variable decision factors (e.g., one factors, three factors) made the decision-making process easier. In addition to subjective ratings, participants' skin conductance and pupillary responses were also collected with GSR sensors and an eye tracker, respectively, during task time.

6. ANALYSIS OF SUBJECTIVE RATINGS

Figure 4 shows the average subjective ratings of task difficulty levels. A Friedman test showed that there was a statistically significant difference among the seven tasks in difficulty levels, $\chi^2(6) = 155.54$, $p < .001$. A post hoc analysis with Wilcoxon signed-rank tests was then conducted, with a Bonferroni correction applied; this resulted in a new significance level set at $p < .007$ ($.05/7 = .007$) for all pairwise differences.

Decision making with/without certain decision factors: The post hoc tests showed that Task 1 was significantly easier than any other tasks except Task 2 ($Z = -2.588$, $p = .010$). Task 2 did not show a statistical difference from Task 1, as we expected from H1. This could be due to the fact that the Bonferroni adjustment is used to avoid any possible Type I errors, but it is known that Bonferroni adjustment overcorrects the alpha level and may cause Type II errors, and hence reduces overall statistical power [Rothman 2010]. Therefore, to avoid any potential Type II errors, we



(a) Setup of the experiment (left), and example signals of GSR (top right) and Eye-Tracker (bottom right).

(b) Screenshot of a decision making task performed in the study (the map was adapted from Google Map ©).

Fig. 3. Setup of the experiment and screenshot of a decision-making task performed in the study.

used a readjusted significance alpha level of .01 to see if we can find the differences we expected. This adjusted alpha level of .01 was calculated by dividing the original alpha level of .05 by 5, based on the fact that we have five conditions to test among tasks. Using this new alpha level of .01, the results showed that Task 1 was slightly easier than Task 2. Task 2 was significantly easier than Task 3 ($Z = -5.102, p < .001$). These results suggested that decision making with/without specific decision factors resulted in differences in the ease of decision making, as we hypothesized (H1).

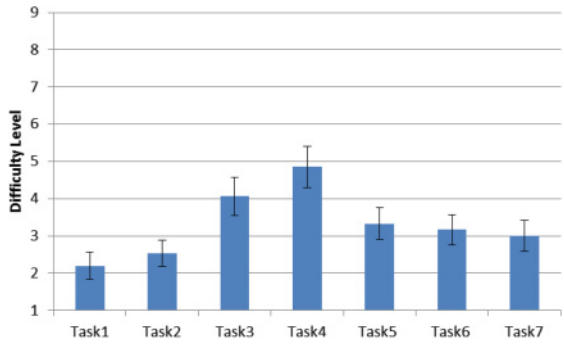


Fig. 4. Subjective ratings of task difficulty levels.

Values of a decision factor: The post hoc tests also showed that Task 2 was statistically significantly easier than Task 6 ($Z = -3.063$, $p = .002$), but there was not a significant difference with Task 7 ($Z = -2.354$, $p = .019$). There was also not a significant difference between Task 6 and Task 7 ($Z = -1.197$, $p = .231$). This means that different values of a specific decision factor affected the ease of the decision-making process. Tasks with different values of a specific factor for various alternatives were statistically easier than were tasks where values of a specific factor were the same, as we hypothesized (H3). However, when values of a specific factor for all alternatives were same, the effect of this factor became less important, and participants mainly considered other factors in their decision making.

Number of certain decision factors: From the post hoc tests, it was found that Task 2 was statistically significantly easier than Task 3 ($Z = -5.102$, $p < .001$) and Task 4 ($Z = -6.074$, $p < .001$). Task 3 was also statistically significantly easier than Task 4 ($Z = -3.859$, $p < .001$). These results suggest that the number of decision factors significantly affected the ease of the decision-making process and that a greater number of decision factors made the decision-making process more difficult. All these findings were in line with our hypotheses (H2).

Different decision factors: The post hoc tests also revealed that Task 2 was statistically significantly different in ease than Task 5 ($Z = -3.568, p < .001$). This suggested that different decision factors affected the ease of decision making differently, as we expected (H4). More specifically, decision making with congestion probability was significantly easier than was decision making with incident probability. The subjective ratings also showed that the least difficult task is Task 1, and the most difficult task is Task 4, as we expected.

7. ANALYSIS OF GSR RESPONSES

GSR responses from 14 participants were analyzed. Figure 5 shows an example of a participant's GSR signals in one task session. As shown in Figure 5, in a decision-making task, there is a 3-second X display before the task begins. GSR responses during both task time and the rest period (X display) are used to analyze task difficulty levels. The GSR data analysis process is divided into following steps: (1) data calibration, (2) signal smoothing, (3) extrema detection, (4) feature encoding, (5) feature significance test, and (6) difficulty level classification.

7.1. GSR Data Analysis

Data calibration: We observed that GSR was highly affected by participants' workload state before task time. To compensate for differences between tasks for each participant, a calibration is applied to each GSR during task time. Because there is no workload

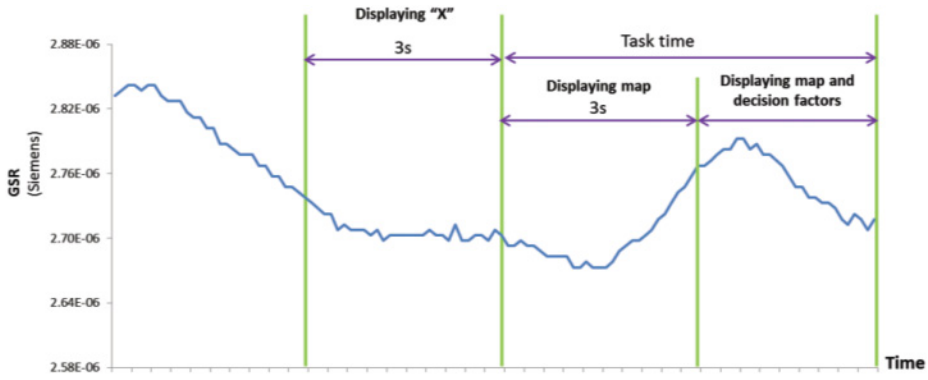


Fig. 5. The time setting of GSR responses in a decision-making task.

during the X display, the average GSR of $\mathcal{G}_X$ during this period is used to calibrate GSR during task time. The calibrated GSR is defined as in the following equation:

$$\mathcal{G}_T = \frac{\mathcal{G}_t - \mathcal{G}_X}{\mathcal{G}_X},$$

where $\mathcal{G}_t$ is the GSR value without calibration during task time, and $\mathcal{G}_T$ is the calibrated GSR of $\mathcal{G}_t$.

Signal smoothing: A Hann window function [Oppenheim and Schafer 2010] is applied to GSR signals to remove noise. The advantage of the Hann window is very low aliasing.

Extrema detection: Extrema-based features can provide various practical benefits through their natural robustness under a variety of practical distortions, their economy of representation, and their computational benefits [Vemulapalli et al. 2013]. Extrema-based features were used in GSR analysis.

We observed that GSR was highly subjective, differing from person to person. To compensate for differences between participants, a calibration is applied to each GSR during task time. Therefore, the smoothed signal is normalized using Z-Normalization to omit subjective differences between various signals before extrema detection. Z-Normalization preserves the range and introduces the dispersion of the series:

$$\mathcal{G}_T(i,j) = \frac{\mathcal{G}_t(i,j) - \mu_{\mathcal{G}_j}}{\sigma_{\mathcal{G}_j}},$$

where $\mu_{\mathcal{G}_j}$ and $\sigma_{\mathcal{G}_j}$ are mean and various GSRs in task j of participant i. After normalization, the extrema points of GSR signals are detected, as shown in Figure 6.

Feature encoding: Both statistical features [Nourbakhsh et al. 2013] and extrema-based features [Healey and Picard 2000] are analyzed. These features include (1) mean of GSR (summation of GSR values over task time divided by task time) $\mu_{\mathcal{G}}$; (2) variance of GSR $\sigma_{\mathcal{G}}$; (3) task time length T_t; (4) number of responses S_f; (5) sum of duration $S_d = \sum S_{di}$; (6) sum of magnitude $S_m = \sum S_{mi}$; and (7) sum of estimated area $S_a = \sum S_{ai}$. S_f, S_d, S_m, and S_a are features of the GSR orienting response [Healey and Picard 2000]. The definition of magnitude S_{mi} and duration S_{di} are defined as shown in Figure 6. The area of response is estimated by $S_{ai} = \frac{1}{2} S_{mi} S_{di}$.

Significance test of GSR features: We performed a one-way ANOVA test with post hoc analysis using a t-test with Bonferroni correction to evaluate task difficulty level discrimination for each feature among the seven tasks. The ANOVA test showed

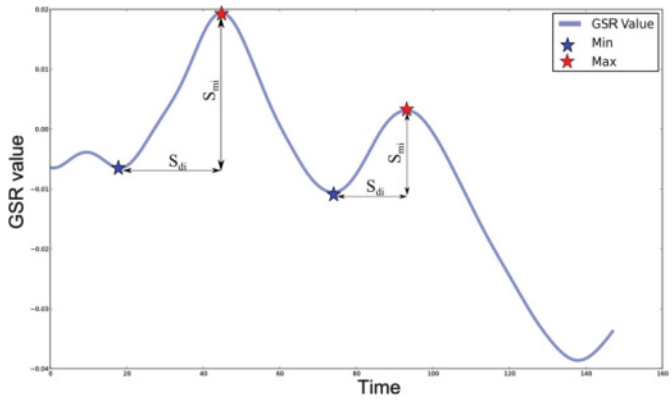


Fig. 6. Extremas and extrema features of GSR.

that features of T_t ($F_{6,287} = 13.889, p < .001$), S_f ($F_{6,287} = 7.956, p < .001$), S_d ($F_{6,287} = 3.41, p = .003$), S_m ($F_{6,287} = 4.302, p < .001$), and S_a ($F_{6,287} = 3.511, p = .002$) showed statistically significant differences among the seven tasks. Post hoc analysis with a t-test was then conducted with a Bonferroni correction (significance level set at $p <$. 01, as discussed in the previous section) for all pairwise differences of five significant features.

The post-hoc tests showed that Task 1 ($t = -3.287, p = .001$) and Task 2 ($t = -3.628$, $p < .001$) had significantly lower values than Task 3 for T_t. These results suggested that decision making with/without specific decision factors resulted in differences in GSR features, as we expected (H1). Furthermore, Task 2 had significantly lower GSR values than did Task 4 for all significant features. Task 3 also had significantly lower GSR values than did Task 4 for all significant features except S_d ($t = -1.557, p = .123$). The results suggested that the number of decision factors significantly affected GSR values and that a greater number of decision factors made GSR values higher, which suggested that the decision-making process was more difficult. All these findings were in line with our hypotheses (H2).

However, there were no significant differences found in GSR values between Task 2 and Task 7, nor between Task 6 and Task 7. This showed that the different values of a specific decision factor did not affect GSR values significantly, as we expected (H3). One possible reason was that the difference between values of a specific decision factor was not high enough to stimulate GSR changes. Also, despite the significant differences in ease found between Task 2 and Task 5 from subjective ratings (as mentioned in Section 6), significant differences between GSR values were not found for all significant features between Task 2 and Task 5, as we expected (H4). This could be because the difference between the two different decision factors of congestion and incident was not high enough to stimulate significant changes in GSR signals.

The comparison of post hoc analysis for subjective ratings and GSR data shows that the results of the post hoc tests for GSR responses were in line with the results of post hoc tests for subjective ratings except that post hoc tests for subjective ratings identified more task pairs with significant differences.

Classification for difficulties of decision making: Support Vector Machine (SVM), Random Forest, C4.5, and Naïve Bayes classifiers were applied to classify decision difficulty levels based on GSR features. Five identified significant features (T_t, S_f, S_d, S_m, S_a) were used to examine two-class classification, where Task 3 and Task 4 were considered high difficulty level tasks while the other tasks were considered low difficulty level tasks based on the subjective ratings presented previously.

Table II. Classification Accuracies with Five GSR-Significant Features for Difficulty Levels

	2-Class Classification			3-Class Classification		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
SVM	0.738	0.119	0.986	0.422	0.095	0.976
RF	0.694	0.321	0.843	0.442	0.500	0.781
C4.5	0.752	0.500	0.852	0.544	0.548	0.848
NB	0.704	0.524	0.776	0.435	0.548	0.748

RF, Random Forest; NB, Naïve Bayes; Acc, Accuracy; Sens, Sensitivity; Spec, Specificity.

We also examined the three-class classification of difficulty levels, where Task 1 and Task 2 were considered low difficulty level tasks, Task 3 and Task 4 were considered high difficulty level tasks, and other tasks were considered middle difficulty level tasks based on subjective ratings discussed earlier. The leave-one-out method was used in the cross-validation. Classification accuracies are shown in Table II. The results show that C4.5 outperforms any other classifiers both for two-class classification (Acc.: 75.2%) and three-class classification (Acc.: 54.4%) cases. It can be observed that GSR can be used to effectively indicate the difficulty levels of decision making to some degree.

7.2. Decision Performance

Decision performance refers to the measurement of whether users choose the most favorable alternative from among multiple alternatives. Various approaches for evaluating decision performance have been proposed [Azizi and Ajirlu 2010; Ray and Sahu 1990]. For example, decision performance is defined as the degree of confirmation to specifications of activities in relation to one or more of their desired values [Ray and Sahu 1990].

Most scholarly quantitative work regarding decision making for travel demand forecasting, transportation planning, and congestion management is based on the concept of utility maximization [Knorring 2003]. To evaluate decision performance, the concepts of *cardinal utility* and *ordinal utility* [Rothbard 1977] are used in this paper. Cardinal utility captures the contribution of the magnitude of decision factors for decisions, whereas ordinal utility captures the contribution of the rank of decision factors for decisions. The cardinal utility function is defined for each alternative with the equation:

$$U_i^c = \sum_{j=1}^{N} w_j u_{ij}^c = \sum_{j=1}^{N} w_j \frac{t_{ij} - t_j^{\min}}{t_j^{\max} - t_j^{\min}},$$

where U_i^c is the ith alternative's cardinal utility, w_j is the weight for the jth decision factor, u_{ij}^c is the jth decision's factor-wise cardinal utility of the ith alternative, t_{ij} is the transformed jth decision factor's value of the ith alternative, and $t_j^{\min}$ and $t_j^{\max}$ are the minimum and maximum of value of t_{ij} of the ith alternative. If $t_j^{\min}$ and $t_j^{\max}$ are the same, u_{ij} is zero. Utility is a measure of satisfaction, and the higher the value, the better. Values of decision factors in each alternative need to be transformed based on their physical meanings. For example, regarding route length, it should be the shorter, the better. Therefore, this value is transformed using the equation: $t_i = v_{max} - v_i$, where t_i is the transformed value, v_{max} is the attribute maximum in alternatives, and v_i is the original attribute value in this alternative. The weight w_j is calculated according to the participant's ranking on decision factors. The rank was collected after all tasks were done. We can regard rank as participants' preferences when they make decisions. If fewer factors are included in a task (i.e., two, three, or four factors), we normalize the

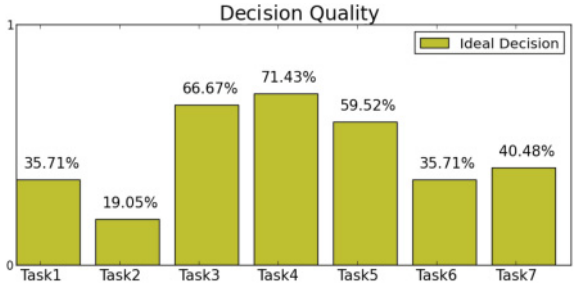


Fig. 7. Decision quality of the seven tasks.

weight by dividing it by the sum of the rank of included factors. This weight definition approach not only reflects differences in decision making among participants, but also considers preferences of decision factors from a participant.

Ordinal utility is calculated according to the rank of factor values among the three routes. If $rank = 1, u^o_{ij} = 1$, otherwise $u^o_{ij} = 0$. The ordinal utility is:

$$U^o_i = \sum_{j=1}^{N} w_j u^o_{ij},$$

where w_j is same as in cardinal utility. The final utility U_i is obtained with the following equation:

$$U_i = (1 - r)\, U^c_i + r U^o_i,$$

where r is the factor controlling the proportion of cardinal utility and ordinal utility in the final utility. Considering the observation that ordinal utility is usually preferred over cardinal utility [Köbberling 2006], r is set in the range of [0.5, 1.0]. It is searched in this range and determined experimentally by considering the decision performance from subjective ratings. r was 0.67 decided experimentally in this study. We assume that the alternative with maximum U_i is the best decision for each participant. This alternative is regarded as the participant-wise ground-truth of a decision. Furthermore, participants may choose different alternatives from preferences as final decisions because they were required to make decisions by considering all decision factors, not preferred decision factors only. Therefore, this participant-wise ground-truth can be used to measure the decision performance of each participant.

Based on the utility definition, the user's decision performance is measured based on (1) computing utility U_i of each alternative; (2) deciding the best alternative that has the highest utility (this can be regarded as the ground-truth of decisions); and (3) comparing the user's decision with the best alternative. If the user's decision matches the best alternative, the user's decision performance is marked as 1. Otherwise, it is marked as 0. This value is defined as the decision quality score. We define decisions that have high scores as high-quality decisions. Figure 7 shows the decision quality of the seven tasks.

Figure 7 shows that Task 1 had a lower decision quality than did any other tasks except Task 2 and Task 6. This suggests that the number of decision factors did help improve decision quality. Task 2 and Task 6 did not have a higher quality than Task 1, as we expected; this may be because there was no apparent preference trend on congestion probability, which was confirmed by participants' ranking during the experiment. This resulted in a decrease of ordinal utility and thus also a decrease in decision quality. When more decision factors were introduced into decision making, the decision quality was increased significantly, which can be seen from the comparison of decision quality for Task 2, Task 3, and Task 4. This result suggests that it

Table III. Classification Accuracies of Decision Quality Based on GSR Features

	GSR Features			GSR Features + Utilities		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
SVM	0.514	0.897	0.080	0.748	0.750	0.746
RF	0.585	0.590	0.580	0.769	0.821	0.710
C4.5	0.643	0.532	0.768	0.799	0.923	0.659
NB	0.578	0.641	0.507	0.731	0.769	0.688

RF, Random Forest; NB, Naïve Bayes; Acc, Accuracy; Sens, Sensitivity; Spec, Specificity.

is necessary to control the number of variable decision factors in decision making in order to get decisions with high quality. Further analyses show that Task 2 had a lower decision quality than Task 5, which means that different types of decision factors affected decision quality differently. It is also shown that Task 2, Task 6, and Task 7 had different decision quality. This suggests that the values of decision factors also affect decision quality significantly.

We examined GSR for quantitative decision quality. Five significant GSR features were used to examine two-class classification of decision quality. The ground-truth of the two-class was set up based on the definition of decision quality mentioned earlier: If the user's decision matches the best alternative having the highest utility, the decision was marked as 1 (of high quality) and the corresponding GSR of task was also marked as 1. Otherwise, the GSR of task was marked as 0. The decision quality classification was examined based on five GSR significant features with/without the utilities of the alternatives using SVM, Random Forest, C4.5, and Naïve Bayes classifiers. Table III shows classification accuracies in two cases. The leave-one-out method was used in the cross-validation. The result demonstrates that the classification based on GSR features plus utilities outperforms the classification based on GSR features only, where C4.5 outperforms any other classifiers. The classification accuracy based on GSR features plus utilities with C4.5 is as high as 79.9%. The results suggest that GSR together with utilities can be used to indicate decision quality effectively in decision making.

8. PUPILLARY ANALYSIS

In this study, the pupil diameter from an eye tracker is investigated to analyze the effects of various decision factors on pupillary response. The average value of left and right pupil diameter is used in the study. We use an approach similar to that used with GSR data analysis (as presented in the previous section) to analyze pupillary data: (1) data calibration, (2) signal smoothing, (3) extrema detection, (4) feature encoding, (5) feature significance test, and (6) difficulty level classification.

Figure 8 shows an example of a pupil diameter signal during task time and its extremas and extrema features after data calibration and signal smoothing. The features of pupillary data used in this study include (1) mean of pupil diameter μ_p, (2) variance of pupil diameter σ_p, (3) task time length T_t^p, (4) number of responses of pupil diameter signal S_f^p; (5) sum of duration of pupil diameter signal $S_d^p = \sum S_{di}^p$; (6) sum of magnitude of pupil diameter signal $S_m^p = \sum S_{mi}^p$; and (7) sum of estimated area of pupil diameter signal $S_a^p = \sum S_{ai}^p$. The area of response of the pupil diameter signal is estimated by $S_{ai}^p = \frac{1}{2} S_{mi}^p S_{di}^p$.

Significance test of pupillary features: We performed a one-way ANOVA test with post hoc analysis using a t-test with Bonferroni correction to evaluate the task difficulty level discrimination of each feature of the pupil diameter signal among the seven tasks. The ANOVA test showed that features of T_t^p ($F_{6,287} = 15.077$, $p < .001$), S_f^p ($F_{6,287} = 10.964$, $p < .001$), and S_d^p ($F_{6,287} = 4.541$, $p < .001$) showed statistically

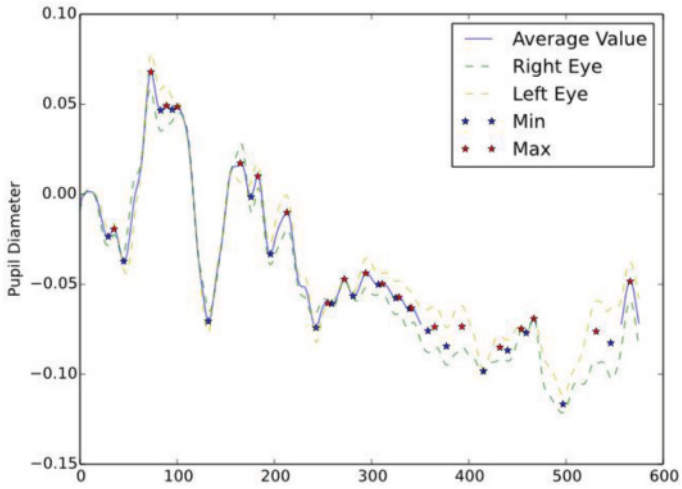


Fig. 8. Extremas and extrema features of pupil diameter data.

Table IV. Classification Accuracies with Three Significant Pupillary Features for Decision Difficulty Levels

	2-Class Classification			3-Class Classification		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
SVM	0.718	0.971	0.083	0.452	0.393	0.829
RF	0.721	0.848	0.405	0.456	0.500	0.714
C4.5	0.629	0.881	0.000	0.442	0.095	0.909
NB	0.725	0.786	0.571	0.493	0.536	0.786

RF, Random Forest; NB, Naïve Bayes; Acc, Accuracy; Sens, Sensitivity; Spec, Specificity.

significant differences among the seven tasks. Post hoc analysis with a t-test was conducted with a Bonferroni correction (significance level set at $p < .017$, which is obtained through dividing .05 by 3 because of three significant features) for all pairwise differences of three significant features. The post hoc tests showed conclusions based on changes of pupillary features to be similar to conclusions based on GSR features as presented in the previous section. For example, the post hoc tests revealed that all other tasks had significantly lower values than Task 4 for all significant features except S_d^p. For S_d^p, all other tasks had significantly lower values than Task 4, except Task 7 ($t = 1.768$, $p = .081$). Furthermore, forT_t^p, Task 1 ($t = -3.287$, $p = .001$), Task 2 ($t = -3.628$, $p < .001$), and Task 6 ($t = 2.449$, $p = .016$) also had significantly lower values than Task 3. ForS_f^p, Task 1 ($t = -2.751$, $p = .007$) and Task 2 ($t = -3.080$, $p = 0.003$) had significantly lower values than Task 3. Similar to the GSR data, the comparison of post hoc analysis for subjective ratings and pupil diameter data showed that the results of the post hoc tests for pupillary responses were in line with the results of post hoc tests for subjective ratings, except that post hoc tests for subjective ratings identified more task pairs with significant differences.

Classification for decision difficulty levels based on pupillary features: We also examined classification of decision difficulty levels based on pupillary features using SVM, Random Forest, C4.5, and Naïve Bayes classifiers. Three identified significant features (T_t^p, S_f^p, and S_d^p) were used to examine two-class and three-class classifications, where the ground-truth of task difficulty levels were the same as those used in the classification based on GSR features (see Table II). The leave-one-out method was used in the cross-validation. The classification accuracies are shown in Table IV.

Table V. Classification Accuracies of Decision Quality Based on Pupillary Features

	Pupillary Features			Pupillary Features + Utilities		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
SVM	0.592	0.885	0.049	0.687	0.890	0.311
RF	0.589	0.738	0.738	0.728	0.822	0.553
C4.5	0.643	0.990	0.000	0.657	0.843	0.311
NB	0.602	0.691	0.437	0.691	0.764	0.553

RF, Random Forest; NB, Naïve Bayes; Acc, Accuracy; Sens, Sensitivity; Spec, Specificity.

Table VI. Classification Accuracies of Decision Quality Based on GSR and Pupillary Features

	Features			Features + Utilities		
	Acc.	Sens.	Spec.	Acc.	Sens.	Spec.
SVM	0.643	0.984	0.010	0.738	0.895	0.447
RF	0.578	0.754	0.252	0.755	0.859	0.563
C4.5	0.626	0.964	0.000	0.772	0.937	0.466
NB	0.595	0.691	0.418	0.653	0.686	0.592

RF, Random Forest; NB, Naïve Bayes; Acc, Accuracy; Sens, Sensitivity; Spec, Specificity.

The results show that Naïve Bayes outperforms any other classifiers both for two-class classification (Acc.: 72.5%) and three-class classification (Acc.: 49.3%) cases. It can be observed that pupil diameter can be used to effectively index the difficulty level of decision making to some degree.

Classification of decision quality based on pupillary features: We examined pupillary features for indexing decision quality. Three significant features of the pupillary signal were used to examine two-class classification of decision quality using SVM, Random Forest, C4.5, and Naïve Bayes classifiers. The ground-truth of decision quality was same as that used in the decision quality classification based on GSR features in Table III. Table V shows classification accuracies in two cases of pupillary features with and without the utilities of the alternatives. The leave-one-out method was used in the cross-validation. The results demonstrate that the classification based on pupillary features plus utilities outperforms the classification based on pupillary features only, and that Random Forest outperforms any other classifiers. The classification accuracy based on pupillary features plus utilities with Random Forest is as high as 72.8%. The results suggest that pupillary features together with utilities can be used to index decision quality effectively in decision making.

Multimodal-based decision-making analysis: In this study, we also combined significant features of GSR and pupillary response together to index decision quality. The decision quality classification was examined based on combined features with/without the utilities of the alternatives using SVM, Random Forest, C4.5, and Naïve Bayes classifiers. Table VI shows classification accuracies in two cases. The leave-one-out method was used in the cross-validation. The result demonstrates that classification of decision quality based on combined features with utilities outperforms the classification without utilities. Furthermore, C4.5 with combined features plus utilities shows the highest classification accuracy (77.2%) in decision quality classification. By comparing Table V and Table VI, it was found that the combined features of GSR and pupillary response show higher accuracies in decision quality classification than pupillary features only. However, in comparing Tables III and VI, the combined features of GSR and pupillary response show lower accuracies in decision quality classification

than GSR features only. Therefore, GSR features with utilities performed the best in indexing decision quality in our study.

9. DISCUSSION

This article investigated MADM with variations of type, number, and values of decision factors. From the results of this study, it was demonstrated that these variations affect both decision quality and the difficulty level of decision making significantly. For example, despite the fact that more decision factors increase the decision difficulty level, they notably helped users improve decision quality. More importantly, it was demonstrated that the effects of these aspects of decision factors on decision making can be measured in real-time to evaluate whether a decision task is at an appropriate difficulty level or whether a decision made by users is of high quality. Therefore, a correlation between physiological signals and measurable decision making was established. This correlation helped users choose and refine decision factors adaptively during the decision-making process, as shown in Figure 1. For example, users may include more decision factors in order to make higher quality decisions.

GSR and pupillary response measurements were analyzed in this study. By comparing Tables II and IV, it was found that GSR features performed better in indexing difficulty levels of decision making than did pupillary features. Similarly, the comparison of Tables III and V showed that GSR features also exhibited better performance in indexing decision quality than did pupillary features. When GSR and pupillary features were combined to index decision quality, we found that the combined features had better performance than pupillary features only but lower performance than GSR features only. However, the combination of features from different channels usually led to better performance in classification than using features from a single channel. One reason for the result of this study could come from pupillary "noise factors," such as luminance changes. Therefore, GSR features with utilities showed the best performance in indexing decision quality in this study.

To incorporate these findings into real-world applications, the user interface for a MADM application needs to include (1) components that allow users to adaptively choose which decision factors are considered in their decision making; and (2) to present the difficulty levels of decision making and decision quality in real-time.

Such a user interface can help users make higher quality decisions. By using physiological sensors such as GSR devices and eye trackers during decision making, the difficulty level and quality of each decision may be measured and displayed in real-time. The real-time feedback of difficulty levels and decision quality allows the user to adjust her selection of factors impacting her decision, in order to balance decision difficulty and quality during the decision-making process.

The proposed framework integrated parsing and interpretation of user physiological information with computational algorithms that, in turn, fed into processes that adapt the interface for decision factors to enhance user performance in decision making. The types of interface adaptations in an intelligent interface that one may consider include (1) the addition or deletion of decision factors, (2) changing the values of decision factors, (3) changing the types of decision factors, and (4) the addition or deletion of signal channels used to measure user's physiological information.

Figure 9 illustrates the loop of using adaptive measurable decision making in an application. In this loop, the adaptive measurable decision making engine is mainly composed of the physiological signal processing component and classifiers for decision difficulty and decision quality, as well as decision factor adaptation. Raw physiological signals from the user are input into the adaptive measurable decision making engine. The decision difficulty levels and decision quality are derived from these signals. If the user is not satisfied with the decision difficulty levels and decision quality, the

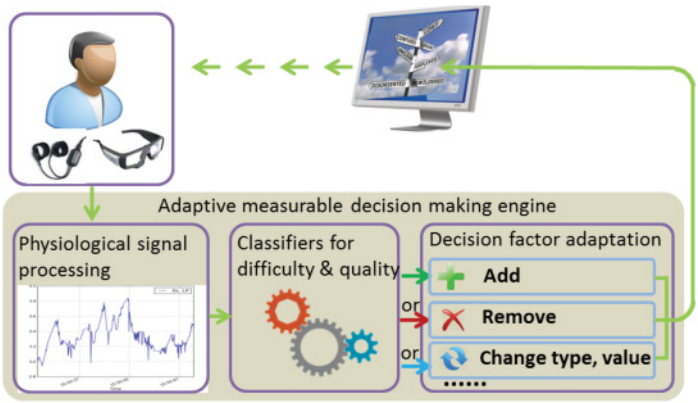


Fig. 9. Diagram of the use of adaptive measurable decision making in an application.

decision factors are refined (e.g., add/remove some decision factors, change types/values of decision factors) and a new decision process is performed based on the updated decision factors. This process is iteratively performed until the user is satisfied with the decision performance.

We expect that our approach will have potential benefits for a broader class of HCI systems based on the diagram shown in Figure 9 (e.g., in wireless computing devices such as cell phones or wearable computers that have built-in cameras and other sensors for physiological signal recordings, and where the user interacts with a decision-support application that displays information on decision factors as a function of the current decision difficulty level and decision quality). In this scenario, decision difficulty levels and decision quality are measured with built-in sensors and updated in real-time based on current decision factors. As a result, such decision-support applications help the user to make higher quality decisions with appropriate difficulty levels.

Furthermore, in most cases, to drive or improve decision making is the ultimate goal of ML-based data analysis [Veropoulos 2001; Kelemen et al. 2002; Krishnamurthy et al. 2009; Helgee 2010]. When an ML approach is used to infer a model from input data, the quality of the model should be judged not from the point of view of how it fits the "true model" but from the point of view of how good the decisions are that one makes based on this model [Bousquet 2005]. Various ML results can be used as decision factors in MADM. Therefore, there is a close connection between ML and decision theory. Previous works [Krishnamurthy et al. 2009; Xu 2009; Zhou and Huang 2012] focus on the direct use of ML results in decision making, such as choosing decision alternatives that have the highest values of ML results as final decisions. If we consider ML results from real ML models as decision factors in the framework of measurable decision making, then various aspects of ML results (e.g., types, numbers, and/or values of ML results) can show close relationships with decision quality and decision difficulty levels, as conventional decision factors do. Following the framework of adaptive measurable decision making, users are aware of which ML models produce ML results for higher decision quality. As a result, ML models can be evaluated not based on ML results directly, but based on decision quality, which is more acceptable to both ML researchers and domain experts. Therefore, this study provided an applicable approach to evaluate ML models for both ML researchers and domain experts. Such an evaluation approach is especially meaningful in real-world applications with big data analytics. In this regard, this study opened a door between ML research and decision making.

In summary, this study showed that physiological signals such as GSR and pupillary response can be used to index decision quality and difficulty levels in a decision-making

process. Human users may modulate the type, number, and/or values of decision factors to adaptively refine decision quality and the difficulty levels of decision making. These findings have at least two benefits in real-world applications:

—The design of intelligent user interfaces for decision-related applications in HCI. In such a user interface, users' physiological signals are collected during decision making. The user interface for a MADM application also needs to include components of choosing various decision factors, as well as presenting decision difficulty levels and decision quality in real-time.
—The evaluation of ML models in ML research areas by employing ML results as decision factors and estimating decision quality in MADM applications.

10. CONCLUSIONS AND FUTURE WORK

This article presented a framework of adaptive measurable decision making. Physiological measurements were used to index difficulty levels and quality of decision making in order to provide real-time feedback on the difficulty and quality of decisions. It was found that various aspects of decision factors (e.g. type, number, and values) affected the decision difficulty level and the quality of decisions. It was also found that multimodal-based measurements performed better in indexing decision quality than pupillary features only, but had lower performance in indexing decision quality than did GSR features only. Finally, we also attempted to formulate guidelines for the design of user interfaces involving decision making.

This article opens a door between decision making and users in order to effectively use various decision factors in decision making. One of our future directions will focus on investigating more effective methods for the measurement of decision making. We believe additional decision analysis approaches can be defined to further improve physiological measurement as an index of decision quality.

AUTHOR STATEMENT

This is the original research in NICTA. We did not submit any part of these research results to any other journals or publications. It bears no relation to our previous publications.

ACKNOWLEDGMENTS

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. The authors thank Benjamin Itzstein, Wei Liu, and Nargess Nourbakhsh for fruitful discussions and help in setting up experiments. The authors also wish to thank all volunteers for contributing their time to experiments.

REFERENCES

M. A. Abdel-Aty, K. M. Vaughn, R. Kitamura, P. P. Jovanis, and F. L. Mannering. 1994. Models of commuters' information use and route choice: Initial results based on Southern California Commuter Route Choice Survey. *Transportation Research Record* 1453, 46–55.

M. Acevedo and J. I. Krueger. 2004. Two egocentric sources of the decision to vote: The voter's illusion and the belief in personal relevance. *Political Psychology* 25, 1, 115–134.

F. S. Azar. 2000. *Multiattribute Decision-Making: Use of Three Scoring Methods to Compare the Performance of Imaging Techniques for Breast Cancer Detection*. University of Pennsylvania.

H. Azizi and S. F. Ajirlu. 2010. Measurement of overall performances of decision-making units using ideal and anti-ideal decision-making units. *Computers and Industrial Engineering* 59, 3, 411–418.

A. Bechara, A. R. Damasio, H. Damasio, and S. W. Anderson. 1994. Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition* 50, 7–15.

A. Bechara, H. Damasio, A. R. Damasio, and G. P. Lee. 1999. Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *Journal of Neuroscience* 19, 5473–5481.

J. R. Bettman, E. J. Johnson, and J. W. Payne. 1990. A componential analysis of cognitive effort in choice. *Organizational Behavior and Human Decision Processes* 45, 1, 111–139.

M. M. Botvinick and Z. B. Rosen. 2009. Anticipation of cognitive demand during decision-making. *Psychological Research* 73, 6, 835–842.

W. Boucsein. 2012. *Electrodermal Activity*, (2nd ed.). Springer.

O. Bousquet. 2005. Machine Learning Thoughts—Decision Making. Retrieved September 20, 2013 from http://ml.typepad.com/machine_learning_thoughts/2005/06/decisionmaking.html.

W. Bruine de Bruin, A. M. Parker, and B. Fischhoff. 2007. Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology* 92, 5, 938–956.

S. Chen, J. Epps, N. Ruiz, and F. Chen. 2011. Eye activity as a measure of human mental effort in HCI. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI'11)*. ACM, New York, NY, 315–318.

M. E. Dawson, A. M. Schell, and C. G. Courtney. 2011. The skin conductance response, anticipation, and decision-making. *Journal of Neuroscience, Psychology, and Economics* 4, 2, 111–116.

Z. Duric et al. 2002. Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction. *Proceedings of the IEEE* 90, 7, 1272–1289.

S. Fiedler and A. Glockner. 2012. The dynamics of decision making in risky choice: An eye-tracking analysis. *Frontiers in Psychology* 3.

B. Figner and R. O. Murphy. 2011. Using skin conductance in judgment and decision making research. In *A Handbook of Process Tracing Methods for Decision Research*. M. Schulte-Mecklenbeck, A. Kuehberger, and R. Ranyard (eds.). Psychology Press, New York, 163–184.

A. M. Franco-Watkins and J. G. Johnson. 2011. Applying the decision moving window to risky choice: Comparison of eye-tracking and mouse-tracing methods. *Judgment and Decision Making* 6, 8, 740–749.

L. A. Gutnik, A. F. Hakimzada, N. A. Yoskowitz, and V. L. Patel. 2006. The role of emotion in decision-making: A cognitive neuroeconomic approach towards understanding sexual risk behavior. *Journal of Biomedical Informatics* 39, 6, 720–736.

J. Healey and R. Picard. 2000. SmartCar: Detecting driver stress. In *Proceedings of the 15th International Conference on Pattern Recognition 2000*. 218–221.

E. A. Helgee. 2010. *Improving Drug Discovery Decision Making Using Machine Learning and Graph Theory in QSAR Modeling*. PhD thesis. University of Gothenburg, Gothenburg, Sweden.

A. Holzinger. 2013. Human-computer interaction and knowledge discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In A. Cuzzocrea, C. Kittl, D. E. Simos, E. Weippl, and L. Xu (eds.). *Availability, Reliability, and Security in Information Systems and HCI*. Lecture Notes in Computer Science. Springer, Berlin, 319–328.

C.-L. Hwang and K. Yoon. 1981. *Multiple Attribute Decision Making: Methods and Applications: A State-of-the-Art Survey*. Springer-Verlag.

E. Á. Juliusson, N. Karlsson, and T. Gärling. 2005. Weighing the past and the future in decision making. *European Journal of Cognitive Psychology* 17, 4, 561–575.

A. Kelemen, Y. Liang, and S. Franklin. 2002. A comparative study of different machine learning approaches for decision making. In *Recent Advances in Simulation, Computational Methods and Soft Computing*. WSEAS Press, 84–141.

M. Kendall. 1962. *Rank Correction Methods* (3rd ed). Hafner, New York.

S.-H. Kim, Z. Dong, H. Xian, B. Upatising, and J. S. Yi. 2012. Does an eye tracker tell the truth about visualizations?: Findings while investigating visualizations for decision making. *IEEE Transactions on Visualization and Computer Graphics* 18, 12, 2421–2430.

J. H. Knorring. 2003. *Basic Human Decision Making: An Analysis of Route Choice Decisions by Long-Haul Truckers*. Bachelor thesis. Princeton University, Princeton, NJ.

V. Köbberling. 2006. Strength of preference and cardinal utility. *Economic Theory* 27, 2, 375–391.

S. Krishnamurthy, G. Thamilarasu, and C. Bauckhage. 2009. MALADY: A machine learning-based autonomous decision-making system for sensor networks. In *Proceedings of International Conference on Computational Science and Engineering 2009*. 93–100.

S. Lertprapai. 2013. Review: Multiple criteria decision making method with applications. *International Mathematical Forum* 8, 7, 347–355.

L. Luo and R. Taib. 2013. Assessing recovery from cognitive load through pen input. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 1353–1358.

K. R. MacCrimon. 1968. *Decision Making among Multiple Attribute Alternatives: A Survey and Consolidated Approach. RAND Memorandum.* The Rand Corporation, Santa Monica, CA.

J. M. McWilliams, C. C. Afendulis, T. G. McGuire, and B. E. Landon. 2011. Complex Medicare advantage choices may overwhelm seniors–especially those with impaired decision making. *Health Affairs (Project Hope)* 30, 9, 1786–1794.

J. V. Neumann. 1944. *Theory of Games and Economic Behavior*. Princeton University Press.

N. Nourbakhsh, Y. Wang, and F. Chen. 2013. GSR and blink features for cognitive load classification. In *Proceedings of INTERACT 2013*.

A. V. Oppenheim and R. W. Schafer. 2010. *Discrete-Time Signal Processing*. Pearson, Upper Saddle River, NJ.

L. K. Payne. 2008. *Skin Conductance Response as a Marker of Intuitive Decision Making in Nursing*. ProQuest.

K. Preuschoff, B. M. 't Hart, and W. Einhäuser. 2011. Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience* 5, 115.

P. K. Ray and S. Sahu. 1990. Productivity measurement through multi-criteria decision making. *Engineering Costs and Production Economics* 20, 2, 151–163.

V. S. Rotenberg and A. B. Vedenyapin. 1985. GSR as reflection of decision-making under conditions of delay. *The Pavlovian Journal of Biological Science: Official Journal of the Pavlovian* 20, 1, 11–14.

M. N. Rothbard. 1977. *Toward a Reconstruction of Utility and Welfare Economics*. Center for Libertarian Studies.

K. J. Rothman. 2010. Curbing type I and type II errors. *European Journal of Epidemiology*, 25(4), 223–224.

S. Sethi-Iyengar, G. Huberman, and W. Jiang. 2004. How much choice is too much? Contributions to 401(k) retirement plans. In *Pension Design and Structure: New Lessons from Behavioral Finance*. Oxford University Press, 83–95.

Y. B. Shin, S. Lee, S. G. Chun, and D. Chung. 2013. A critical review of popular multi-criteria decision making methodologies. *Issues in Information Systems* 14, 1, 358–365.

P. J. Smith, N. D. Geddes, and R. Beatty. 2009. Human-centered design of decision-support systems. In A. Sears and J. A. Jacko (eds). *Human-Computer Interaction: Design Issues, Solutions, and Applications*. CRC Press.

K. E. Stanovich and R. F. West. 2008. On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology* 94, 4, 672–695.

C. Stickel, M. Ebner, S. Steinbach-Nordmann, G. Searle, and A. Holzinger. 2009. Emotion detection: Application of the valence arousal space for rapid biological usability testing to enhance universal access. In C. Stephanidis (ed). *Universal Access in Human-Computer Interaction. Addressing Diversity*. Lecture Notes in Computer Science. Springer, Berlin, 615–624.

P. K. Vemulapalli, V. Monga, and S. N. Brennan. 2013. Robust extrema features for time-series data analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 6, 1464–1479.

K. Veropoulos. 2001. *Machine Learning Approaches to Medical Decision Making*. PhD thesis. University of Bristol, Bristol, UK.

W. Wang, Z. Li, Y. Wang, and F. Chen. 2013. Indexing cognitive workload based on pupillary response under luminance and emotional changes. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces (IUI'13)*. ACM, New York, NY, 247–256.

XEROX. 2013. Xerox Researchers Hit The Streets to Help Unclog Highways, Reduce Pollution and Find You a Parking Space. Retrieved September 20, 2013 from http://news.xerox.com/news/Xerox-researchers-create-new-transportation-applications.

H. Xu. 2009. *Robust Decision Making and Its Applications in Machine Learning*. PhD thesis. McGill University, Montréal, Canada.

J. Xu et al. 2011. Pupillary response-based cognitive workload index under luminance and emotional changes. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. ACM, New York, NY, 1627–1632.

J.-B. Yang and P. Sen. 1994. A general multi-level evaluation process for hybrid MADM with uncertainty. *IEEE Transactions on Systems, Man and Cybernetics* 24, 10, 1458–1473.

J.-B. Yang and D.-L. Xu. 2013. Evidential reasoning rule for evidence combination. *Artificial Intelligence* 205, 1–29.

Q. Zhou and Z. Huang. 2012. A decision-making method using knowledge-based machine learning. In *Proceedings of International Conference on Computer Science and Electronics Engineering 2012*. 616–620.

Received December 2013; revised September 2014; accepted September 2014