10-16-2017

# Visual Analytics Evaluation Based on Judgment Analysis Theory

Mosab Khayat
*Purdue University - ECE*, mkhayat@purdue.edu

Arif Ghafoor
*Purdue University*, ghafoor@purdue.edu

**Authors:** Mosab Khayat and Arif Ghafoor

**Institution:** Purdue University – ECE

**Email:** mkhayat@purdue.edu, ghafoor@purdue.edu

**Title:** Visual Analytics Evaluation Based on Judgment Analysis Theory

**Date of report:** 10-16-2017

**Keywords:** Evaluation, Quantitative, Idiographic statistics, Judgment Analysis, Lens Model, Situation Awareness, Insights.

**Abstract**: In this paper we propose a framework to quantitatively evaluate user awareness and the level of support that visual analytics decision support systems (VADS) provide. For the framework, which has a theoretical underpinning from the field of judgement analysis, we propose a model for VADS system. The framework bridges the gap between judgment analysis and VADS evaluation by conceptually connecting judgment analysis concepts to visual analytic. The proposed approach offers an insights based evaluation to measure the importance and the utility of the insights. We propose to model insights and user findings as random variables that parametrize user decisions. The mixed methodology used in our framework has the potential to study user decision process in real situations while producing results that can be generalized. Our contributions in this work appear in the modeling of VADS system and the evaluation framework we propose which quantifies situation awareness. Other advantages include evaluating collaboration and analyzing joint decisions. Some limitations of the framework are also discussed including the requirement of large testing data.

# Visual Analytics Evaluation Based on Judgment Analysis Theory

Mosab Khayat, *Student Member, IEEE*, and Arif Ghafoor, *Fellow, IEEE*

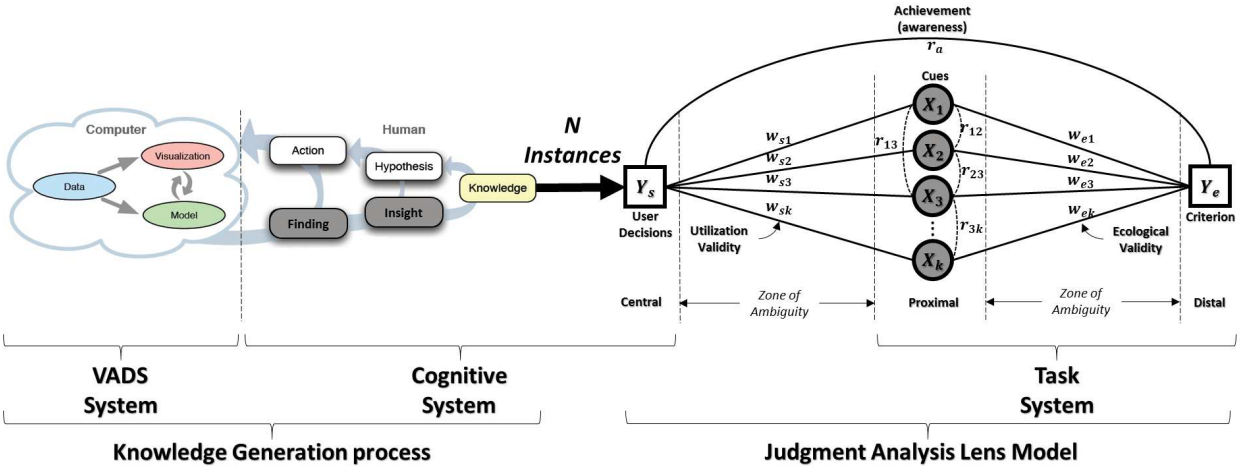


Fig. 1: The proposed relationship between visual analytics and judgment analysis. The left part of the figure shows the knowledge generation model [33]. The right part shows the lens model for analyzing user decisions $Y_s$ collected over $N$ instances. The user utilizes a set of cues $X_1$,$X_2$,...,$X_k$ which is found during the analysis to render the final decisions. When evaluating these decisions, we need criterion or ground truth $Y_e$ to validate the importance of the set of cues used by the user to answer the decision task.

**Abstract**—In this paper we propose a framework to quantitatively evaluate user awareness and the level of support that visual analytics decision support systems (VADS) provide. The framework bridges the gap between judgment analysis and VADS evaluation by conceptually connecting judgment analysis concepts to visual analytic. The proposed approach offers an insights based evaluation to measure the importance and the utility of the insights. We propose to model insights and user findings as random variables that parametrize user decisions. The mixed methodology used in our framework has the potential to study user decision process in real situations while producing results that can be generalized. The proposed framework can also be used for evaluating collaboration and analyzing joint decisions. Some limitations of the framework are also discussed including the requirement of large testing data.

**Index Terms**—Evaluation, Quantitative, Idiographic statistics, Judgment Analysis, Lens Model, Situation Awareness, Insights.

## 1 Introduction

Visual analytics is an emerging problem solving methodology that combines the power of humans and machines to gain a deeper understanding of complex problems to reach optimal solutions that are not achievable by humans or machines working independently. This intertwined environment is created by designing interactive tools that employ automatic data analysis and visualization techniques to improve the awareness of expert users and support their decisions in their domains. The importance of having metrics and formal methodologies to evaluate user awareness has been noted by the VAST community [35]. The awareness of a situation is the key to making the right decisions that visual analytics systems seek. There is a clear connection between visual analytics as a decision support tool and decision theories that study decision related issues.

One can view the visual analytics process, which includes both human and machine involvement, as a black box that takes raw data as an input and provides decisions about a context as an output. Decision theories in general apply one of the following functions to such decision generator black boxes: prescription, prediction, description and explanation [5]. Prescription theories, such as the Analytical Hierarchical Process (AHP) [31], provide guidelines for the black box components to reach optimum decisions in complex contexts. Prediction theories, such as judgment analysis, aims to model the black box outcomes in order to predict its decisions for future situations. Judgment analysis also is an example of description theories which try to open the black box to provide an understanding of how it generates the decisions and what factors control this generation. Finally, explanation theories, such as Attribution Theory [14], ask why the black box makes these decisions. From an evaluation point of view, we want to understand how the black box generates decisions and how good this generation policy is. In other words, our objective is description functionality.

In addition to the intent of the theory, we seek to create a framework that is applicable in a realistic setup yet generates results that can be generalized. Using qualitative methodology as the basis for the framework allows us to apply it in a realistic setup and understand a specific case accurately, but prevents us from generalizing the findings to other cases. In contrast, using quantitative methodology in the framework allows us to generalize findings but may not be applicable in many realistic scenarios entailing uncontrollable variables. When selecting a theory to evaluate visual analytics, it is important to insure that it satisfies the desired intent and the desired scope of evaluation.

Considering these requirements, we observe that judgment analysis theory [5] can be applicable to evaluate visual analytics decision support tools. Its mixed methodological nature provides adaption to many realistic scenarios as well as a way to generalize its findings. The theory follows an idiographic philosophy which is appropriate to describe many visual analytics evaluation practices.

The novelty of our proposal is that we model the main user processes

• *E-mails: (mkhayat, ghafoor) @purdue.edu.*

usually relied upon when evaluating visual analytics systems. We then propose an evaluation framework to evaluate one of the processes in the model. Our framework basically describes the policy that is followed by users' cognitive systems when generating decisions by regressing user decisions to the findings and insights found. This policy description can then be compared to criterion to find the achievement, a metric that can be used to evaluate situation awareness.

This paper is organized as follows: Section 2 reviews some of the related works and provides some background information. Section 3 present our evaluation framework. at the beginning, we formulate the problem of evaluating situation awareness. We then discuss the proposed framework by providing conceptual connections between the components of visual analytics and judgment analysis studies. Section 4 extend the discussion to include additional evaluation metrics such as learning rates and joint decisions. A comparison between our framework and existing related frameworks is provided in section 5 and finally, we present a conclusion and possible future works in section 6.

## 2 Background And Related Works

Understanding user analysis, reasoning and decision making processes have been the focus of many evaluation studies. Existing frameworks such as MILC [38], and insight based evaluation [34] have been used to unfold these user processes. Some studies extend insights based evaluation to understand how users reached their insights. One of these studies was conducted by Guo *et al.* [16] who proposed an evaluation pipeline to enhance traditional insight based evaluation. Their pipeline incorporates quantitative analysis that shows the correlation between different interaction patterns and insights. Our proposal also applies correlation analysis to analyze the generated insights. However, unlike the Guo *et al.* study, we propose correlating insights with users' decisions to capture their decision policies instead of analysis strategies.

The WeightLifter study [27] intersects with our proposed framework. Both works use similar concepts and formalization but target different problems. In [27], a system is used to inform the decision maker about the consequences of changing the policy (the weight vector) on the decision. The system allows exploring several decisions with respect to different cues' weights (or criteria weights as they are called in [27]) that are adjustable by the user. The problem we target is different. As an evaluation approach, our study aims to find the weight vector (decision policy) that optimally describes the variety of decisions that are made by the user. We propose that by collecting user's decisions over multiple judgment profiles, we can use a correlational analysis to find the weight that the user's cognitive system employs.

Applying judgment analysis theory to validate the role of a visualization system in a judgment context has been proposed by Miller *et al.* [23]. In this work, the authors study how to aid decision makers to reduce the bias that can appear in their decision by narrowing the decision space using a visualization system. The focus of that study was to compare an unaided decision maker with a visually aided decision maker. Our models and formulation provide a more holistic framework that is capable of capturing other decision analysis factors besides decision bias studied by Miller *et al.*.

Five evaluation metrics proposed by Scholtz [35] which need to be considered when evaluating visual analytics include: 1. the level of situation awareness; 2. the degree of collaboration; 3. the capability of enabling interaction 4. the creativity of the user interface; and 5. the utility in the system that improves user performance. Adagha *et al.* [1] use these metrics to categorize visual analytics decision support tools. Our framework targets evaluating such decision support tools by redefining some of these metrics and evaluate them quantitatively.

### 2.1 Judgment Analysis

Judgment analysis is a methodology developed to capture and evaluate the policy of a decision maker. The main idea was proposed in the mid 20th century by the Austrian psychologist Egon Brunswik [3] who pointed out that it is possible to divide an environment or a situation that an organism interacts with into distal and proximal parts. The organism observes a set of cues in the environment, the proximal part, which probabilistically relates to the distal part that the organism seeks. It then acts based on what it observes in a probabilistic manner and desires to match the probabilistic behavior of the distal part. Brunswik argues that it is suitable to use correlational statistics to analyze this environmental setup by capturing the variability of its components using samples collected over multiple observations. This idea is then applied to social judgment studies and eventually given the name judgment analysis [5].

The goal of judgment analysis studies is to find a relationship between a human judgment and a set of cues that represents the environment as well as a relationship between these cues and the correct judgments (the ground truth) and ultimately, compare both relationships to determine how skillful a decision maker is. Such relationships are called judgment policy and ecological validity and they can be captured by using regression analysis on samples that can be collected over multiple judgment profiles or instances.

### 2.2 Situation Awareness

The definition of situation awareness as an evaluation metric can be derived from [10] as the degree of understanding current facts and the accuracy of predicting possible future events. It is a metric that has been used traditionally to evaluate pilots. However, Scholtz [35] proposes using it as an evaluation metric for visual analytics users.

Three levels of situation awareness have been proposed by Endsley [10]. The first level is the perception of elements in the current situation. The elements are pieces of information that represent the environment and should be perceived by the user. A higher level of awareness is subsequently obtained by connecting the information to create a bigger picture of the current situation. Creating this bigger picture correctly determines the second level of awareness: the comprehension of the current situation. The last level of awareness is the projection of future status and is obtained when the information is perceived and comprehended to a level that enables accurate prediction to possible future events or situations.

Several methods have been used to evaluate the level of situation awareness [10] [35]. Performance-based methods evaluate the awareness by evaluating the correctness of the decisions made by the decision makers. One of the shortcomings of this method is that even with perfect awareness, a human can render a wrong decision. The second method is knowledge-based which evaluates the awareness by analyzing data that are collected using verbal methods like Think-aloud. The problem associated with such methods is the adverse impact that can be observed in the subjects performance. A third method uses the decision makers' subjective opinions to evaluate their own awareness. It is clear, however, that this method is not accurate as it is possible for a decision maker to believe he has a certain level of awareness while the reality is different. This is due to a limited knowledge of the environment that a decision maker may have.

A common method used to evaluate awareness is the Situation Awareness Global Assessment Technique (SAGAT) [10]. In this method, a qualitative study is conducted with domain experts to derive a set of questions to analytically assist the level of awareness. The decision maker is then placed in a simulation that mimics the real situation and asked at random times to answer some questions selected randomly from the questions developed in the qualitative study. An example of using this method to evaluate the awareness of a user who interacts with human-robot interface is done by Scholtz *et al.* [36] [37].

Judgment analysis theory and its conceptual device, the lens model, can be used to model and evaluate the awareness of an operator working in a human-machine system as proposed by Kirlik and Strauss [20] [42]. They use the expanded lens model (discussed in section 3.2.3) to evaluate situation awareness and note that awareness can be adversely affected by both the unreliability of the operator and the unreliability of the system that presents the information. To accurately measure situation awareness, both parts need to be considered.

By linking the arguments proposed in [35] and [20], we observe that it is possible to evaluate visual analytics by evaluating the change in situation awareness of users which can be captured quantitatively using judgment analysis theory. Similar to the Kirlik and Strauss studies [20] [42], the framework we propose uses the achievement of VADS users as a metric to evaluate the awareness. However, we use

different methodology when applying judgment analysis theory to meet the requirements of evaluating visual analytics user awareness.

### 2.3 Research Methodologies

Humanity uses multiple possible approaches to study and explain observed phenomena. These approaches build theories that expand our knowledge about these phenomena. Traditionally, experiments and studies are classified based on their selected approaches into three main classes: quantitative, qualitative and mixed studies. In this section we will briefly discuss the philosophies and technicalities that separate these classes from one another and which methodology applies to our study. For more extensive details, we refer the reader to [8] [9] [11] . Readers are also referred to Carpendale's paper [4] for an explanation of these approaches from a visualization evaluation perspective.

The German philosopher, Wilhelm Windelband, introduced two terms to describe the tendencies of expanding knowledge [45]. Nomothetic methodology uncovers what is true in general by analyzing observations as a whole and dealing with the differences among them as errors. Idiographic methodology, on the other hand, uncovers what is true in a particular situation by studying cases individually and considering the differences among them as the target of the study. For a more detailed discussion, a recent study [30] has been done to trace the historical differences between the two methodologies.

Quantitative studies aim to build theories that can be generalized, so quantitative studies usually follow the nomothetic methodology. Inferential statistical tests are frequently used in these studies to analyze data and either conform or reject hypotheses that assume something about the population that is represented by experiment subjects. In most cases, researchers are interested in finding causation which requires controlled experiments, i.e., experiments that proceed by changing a single variable and fixing others to find cause and effect relationship. However, this is hard to achieve in a realistic setup which has many uncontrolled variables.

Unlike quantitative studies, qualitative studies seek comprehension of a specific situation instead of an explanation of what is true in general. The aim in these studies is not to generalize findings but to build theories that describe experiment findings in particular. In other words, it applies the idiographic methodology. Transferability, or generalizing qualitative studies results to other cases, is sometimes taken into consideration but is not the objective of qualitative studies, and it is hardly achievable due to the usage of qualitative data. Because of these properties, qualitative studies can be applied in realistic setup to describe specific situations.

Mixed studies combine quantitative and qualitative approaches in order to benefit from the properties of each. The first form of mixed studies is to conduct two consecutive studies, one qualitative and the other quantitative, to support one another. This form can be divided into exploratory and explanatory studies according to which type of experiment to begin with. exploratory conducts an initial qualitative inquiry to gain information that is then used in the quantitative study. In contrast, explanatory studies perform a quantitative study to confirm or reject hypotheses followed by a qualitative study that explains factors which lead to the results. The other form of mixed studies is when the researcher merges both quantitative and qualitative approaches in the same study which is called the embedded mixed study.

Our framework applies judgment analysis to evaluate a particular human subject. The methodology used in these studies is an idiographic mixed methodology. The intention of the framework is to unfold the decision policy which is a unique attribute to the particular user we evaluate. The framework extends insight-based evaluation which can be used to capture insights and findings qualitatively and then applies quantitative analysis to find a relationship between user decisions and the captured information.

## 3 Proposed Framework

It is possible to evaluate Visual Analytics Decision Support tools (VADS) holistically using judgment analysis methodology. The goal of visual analytics in general is to support making the right decisions. This support is achieved using a technology that provides automated statistical analysis methods and interactive visual representation that creates a dialog between a human and a machine to increase the amount of knowledge about the environment under analysis. A description for visual analytics process is proposed by Sacha *et al.* in their Knowledge Generation Model [33]. The model shows how raw data is transformed into pieces of knowledge through visual analytics components. This process continues until the user reduces the uncertainty of the decision task to a level that allows making the final decision. Final decisions are nothing but judgments that are made based on a policy employed in a user's cognitive system that utilizes found pieces of information.

### 3.1 Formulation Of Awareness Evaluation

In this section, we propose a mathematical formulation for the problem of evaluating the awareness and the solution proposed in our framework. We also show the difference between measuring the awareness in our framework and existing performance-based evaluation methodology. Three main components are considered in this formulation: decisions, cues and ground truth. Readers are encouraged to return to Figure 1 during the discussion for an overall picture.

Evaluating the performance of users interacting with a software system has been widely applied in visualization and visual analytics. Common metrics used in these types of studies are the error counts or, in other words, the accuracy of users in answering given tasks. In VADS evaluation, these tasks can be assimilated as decision problems that require interaction to extract information and confirm hypotheses. When evaluating a tool to measure its impact on users performance objectively, nomothetic statistics are applied. Lets consider an experiment with a single task. To prove that a particular design is better than another in assisting users to answer that task in general, we evaluate both tools with $\boldsymbol{n}$ users and collect their answers for that task. This creates two n-dimensional vectors corresponding to users' decisions with the two tools (or design choices). The two vector elements are then compared against a scalar called the ground truth that is known to the researcher to find two new n-dimensional vectors representing the score of subjects in matching that ground truth. These two vectors are then considered as a two sample distribution drawn from a targeted user population with two sample means and two sample standard deviations equal to the means and the standard deviations of each vector respectively. Assuming normality, a statistical test such as a t-test (in the case of comparing two sample distributions) or an f-test (in the case of multiple tools or multiple sample distributions) is then conducted to prove statistically that the same behavior (i.e. improvement in performance when using one tool over another) is true in general for the population of targeted users and not observed due to random errors. [1]

One can observe that in previous types of studies, we treat multiple subject behaviors with a particular treatment as a single behavior with errors in different instances (i.e. the mean and the standard deviation of the performance scores in the n-dimension vectors). However, this nomothetic methodology does not take into account the effect of the variability of the ground truth on the behavior of users. The ground truth used in performance evaluation studies is a scalar that represents an instance of a situation. A situation can be represented by an N-dimensional vector ($\mathbf{Y_e}$) representing $\boldsymbol{N}$ variabilities or instances of the ground truth in that context. To be able to measure the awareness of individual users, we need to measure their decision accuracy in multiple instances of the situation. This suggests that awareness should be measured in the individual level before being aggregated nomothetically. In other words, we need to collect $\boldsymbol{N}$ decisions for an individual user to measure his/her awareness against the N-dimensional vector of situation. This argument is derived from the version of Brunswik [3] who initiates the idea of idiographic statistics.

A direct comparison (e.g. correlation analysis) between the N-dimensional vector of a user's decisions ($\mathbf{Y_s}$) and the N-dimensional vector of the ground truth ($\mathbf{Y_e}$) is a useful quantitative measure to evaluate the awareness of that user in a particular context. Having $\boldsymbol{n}$ users, each with a N-dimensional vector of decision, allows us to evaluate the

[1] This study design is called repeated-measures or within-subjects design. Note that a similar nomothetic intention can be seen in between-subjects design.

awareness of each one of them separately. But merely quantitative values for their awareness level might not be what researchers ultimately seek. More information to diagnose the awareness levels of the users to explain their behavior differences in a particular situation is desired. To achieve that goal, we need to define $\boldsymbol{k}$ pieces of information that have been observed by the users when they make their decisions about the situation instances. These pieces of information are called cues and can be represented for an individual user by a $(N \times k)$ matrix-like table $\boldsymbol{X}$. The $i$th row in $\boldsymbol{X}$ represents a set of $\boldsymbol{k}$ cues that is observable by that individual user when making his/her $i$th decision. The user might actually observe a subset of the $\boldsymbol{k}$ cues so the rows of $\boldsymbol{X}$ might contain empty or null values. Defining the cues and their variability in every $\boldsymbol{N}$ situation instant allows us to find a relationship between them and the user's decision vector. This enables us to model the user decision process in that situation with a function $f$ that parametrizes the user's decision using the $\boldsymbol{k}$ cues as parameters. This function $f$ is called the decision policy of the user. Similarly, we can model the situation using a function $\bar{f}$ which again uses the cues as parameters. The function $\bar{f}$ can validate the importance of the cues for a particular situation. Decomposing the awareness level into a measure that uses these two functions provides a better diagnosis of the factors that affect users' awareness level.

In summary, we need the following components to evaluate the awareness of a user: a vector of N decisions ($\mathbf{Y_s}$), a situation vector with N ground truth ($\mathbf{Y_e}$) and an $N \times k$ table of cues ($\boldsymbol{X}$). These components can be seen in Figure 1. Finding relationships among the vectors and the columns of $\boldsymbol{X}$ enables us to measure the awareness of that user and the reason for it. The following sections provide more thorough discussion on how to define these components in VADS evaluation.

### 3.2 Evaluating The Awareness Of VADS Users By Employing Judgment Analysis Methodology

Figure 1 show our view of the relationship between visual analytics and judgment analysis. We use the Knowledge Generation Model [33] to show the process of generating the information used to make decisions. By testing the VADS system over $\boldsymbol{N}$ decision tasks (formally called judgment profiles), we obtain the vector of user decisions $\mathbf{Y_s}$. We also obtain a set of insights and findings $\{\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_k}\}$ that the user generates and finds during the analysis phase. We model these pieces of information (cues) as random variables to note that the same insights or findings can have variability in their values as we will discuss in the next section. This allows us to consider the found values for these pieces of information over multiple judgment profiles as samples drawn from these random variables. These values construct table $\boldsymbol{X}$.

Defining the set of cues needed for a particular decision task is a major part in any judgment analysis studies and is usually done by qualitative studies such as interviewing experts. This is usually done prior to quantitatively analyzing judgments and, thus, we can consider judgment analysis studies as exploratory mixed studies. However, for some complex tasks, researchers might not be able to determine information directly from the proximal part. VADS usually targets such complex environments which require deep analysis to extract information. The analysis phase starts with the information forging cycle [28]. Combining this with user reasoning capabilities creates a set of information that either is extracted directly as a result of exploration or is generated from user comprehension of this extracted information. So, unlike judgment analysis studies, the cues in VADS evaluation studies can be determined by the researcher during the experiment by combining the information that is generated or found by the cognitive system of the expert experiment subjects in every judgment profile. Several methods have been used to capture information like insights from a user's cognitive system including think-aloud protocol [47] and using the diary method [34]. These methods allow us to extract the cues from the cognitive system of users.

The second part of the environment, based on Brunswik's thoughts, is the distal part which contains the unknown truth or more precisely the correct decision for tasks. Similar to user decisions, ground truth for all judgment profiles constructs the vector $\mathbf{Y_e}$ which can be used as criterion for user decisions correctness. As an illustration, lets consider a user of jigsaw system [13] who has the task to discover suspicious terrorists from analyzing multiple text documents. The user might find cues such as relationships between entities which allow her to judge who is involved in terrorist activities. (see [12] submitted in VAST contest 2007 [15]). The ground truth for that contest can be considered as the criterion that the analyst seeks.

Criterion can be defined in several ways. In traditional judgment analysis studies, researchers either observe it directly from the environment or by interviewing domain experts. It is also common to use synthetic criterion to test a decision maker in the lab, however, this should be performed carefully to ensure an accepted level of ecological validity. The same arguments can be found in visualization and visual analytics literature. For instance, we can see works such as the one proposed by Whiting et al. [44] who use realistic data sets and criterion to validate the results of experiments that test a user's analytical reasoning. In that study, the authors implement a system that assists generating synthetic, yet realistic, data with an embedded criterion (or ground truth) and discuss how to validate such a data generation process. Their contribution has been applied in multiple VAST challenges which can be seen as examples of synthetic environments with distal parts containing synthetic criterion.

After defining the vector of judgment $\mathbf{Y_s}$, the vector of criterion $\mathbf{Y_e}$ and the table of cues $\boldsymbol{X}$, we can conduct correlation analysis to capture relationships between these components. The relationship between $\mathbf{Y_s}$ and the columns of $\mathbf{X}$ describes the utilization of the cues that is employed by the user's cognitive system when making decisions. Using methods such as multiple regression analysis allows us to define the function $f$ (the policy) as an additive linear function that utilizes the cues with a weight vector $\mathbf{W_s} = \mathbf{[W_{s1}, W_{s2}, ..., W_{sk}]}$ that optimally describes the user decisions. Similarly, we can find a relationship between $\mathbf{Y_e}$ and the columns of $\mathbf{X}$ which is called the ecological validity. This name is given to this relationship because it validates the cues relationship to the desired criterion optimally. This relationship can be computed in the same way as the decision policy. The weight vector $\mathbf{W_e} = \mathbf{[W_{e1}, W_{e2}, ..., W_{ek}]}$ indicates the validity of the cues.

Another correlation can be computed between the decision vector $\mathbf{Y_s}$ and the criterion vector $\mathbf{Y_e}$. This correlation is called the achievement $r_a$ and it is used to measure the level of situation awareness quantitatively. A decomposition of this quantity is proposed in judgment analysis literature to diagnose its value with respect to user policy and ecological validity. We discuss how to calculate this value and how to decompose it in section 3.2.3.

Based on previous discussion, one can observe three roles for VADS tools to support user decision making and optimize cues utilization. These roles are described in Figure 2. The first role is to assist the user in finding cues that might not be observable without the tool. An example of such cues is the insights that can be reached by understanding complex patterns or relationships in the data. The second role is signifying substantial cues to influence their utilization. last role is omitting irrelevant cues to prevent misleading.

#### 3.2.1 Coding Schema For Cues And Judgments

Judgment analysis studies model environment components (judgments, cues and criterion) in a particular situation as random variables. To use idiographic statistics, we need to collect samples from these random variables over multiple judgment profiles. the sampling enables studying the relationship between the components considering possible differences among situation instances. Correlational statistics such as regression analysis are used to find statistical evidence about the relationship between the components. But such techniques require having these environmental components in a quantitative form. In this section, we discuss a schema to code the set of cues (i.e. user findings and insights) using the levels of measurement proposed by Stevens [39]). The same schema can be used to code the criterion and the decisions.

The schema classifies cues based on their suitability to measuring levels. Some cues can satisfy the requirement of quantitative levels such as interval or ratio scale, whereas others can only satisfy the basic measurement requirement (i.e. mutual exclusiveness and exhaustiveness) which restricts them to qualitative measurements (i.e. categorical and

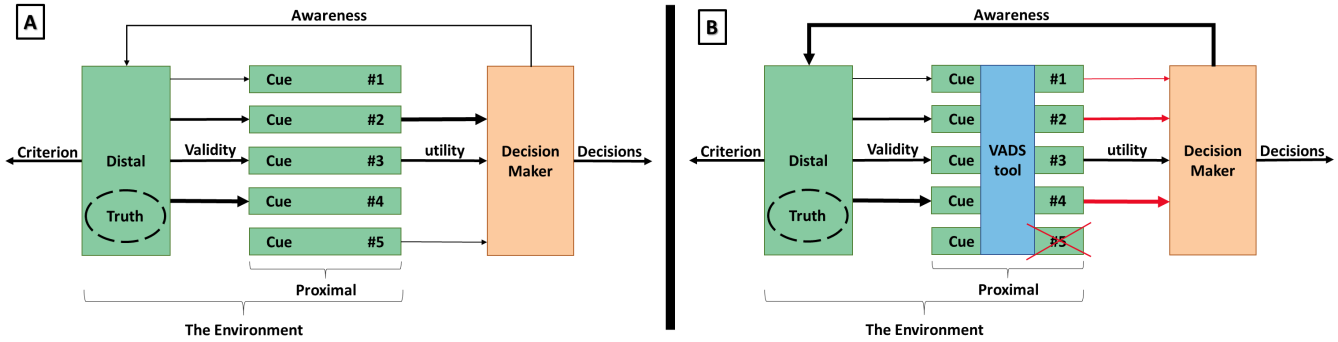


Fig. 2: The role of VADS tools using judgment analysis concepts. Wider arrows = larger weights. No arrow from a cue to the decision maker means the cue is either not utilized. No arrow from the distal to a cue means the cue is not related to the decision task. Part (A) on the left shows the decision analysis without a VADS tool. Differences between user utilization of the cues and the ecological validity of these cues exist due to a limitation in decision maker awareness of the situation. Part (B) on the right shows the decision analysis with the usage of a VADS tool. The tool aim at maximizing the decision maker's awareness by improving her cues utilization. This is done by assisting the decision maker to find unobserved cues (cues #1 and #4), signifying important cues (cue #2) or omitting irrelevant cues (cue #5).

ordinal levels). Considering the suitable measuring level, we classify the cues into quantitative and qualitative cues.

A quantitative cue satisfies the condition of interval or ratio-scale levels. Besides satisfying basic conditions of measuring which allows categorizing the possible values of the cue, it is also possible to quantify a distance between these categories. An example of such types of cues is the average temperature of a city that a user can find by exploring the data. This can be represented by numeric values such as $50°F$, $67°F$, ...*etc*. A cue of this type is represented by quantitative values and, thus, does not need coding.

Some cues do not contain the information needed to be quantitatively measured. It is not possible to define a distance between the possibilities that a cue of this type can take. A qualitative cue is a cue that only satisfies the basic conditions of measuring. It can have a range of possible values that can be categorized into a categorical system without knowing a distance between these categories. Some examples of such types of cues are insights that can be reached by the user such as "There is relationship/no relationship between X and Y" or "The side effect of the medicine is small/medium/large". These facts are not quantitative, but they can still be coded in a way that permits statistical analysis. For instance, a coding technique that is commonly used in regression analysis to handle categorical variables is the dummy coding [17].

Creating a suitable categorical system to measure qualitative cues depends on the skills and knowledge of the researcher who conducts the evaluation study. Researchers might need to consult domain experts to assist in creating the right categorical system for some cases. Other guidelines that are beneficial in this topic have been proposed in the Grounded Theory [6].

A special case of coding can be used to study the effect of cue existence. In this case, effect of observing/not observing a particular cue on the policy of the user is targeted. For instance, it is possible to study the effect of observing a particular pattern or relationship between two entities in the data. To perform such analysis, researchers can evaluate the decision of the user in multiple judgment profiles with a variability of this cue existence which can be treated as a dichotomous categorical variable and coded as categorical cues.

### 3.2.2 Capturing The Decision Policy Of a User

Policy capturing is the procedure to mathematically model (parameterize) judgments as a dependent variable which depends on the utilized cues. It is the process of finding the optimum weight vector $\mathbf{W_s}$. The same process can be used to find the optimum ecological validity vector $\mathbf{W_e}$. Common procedures used in the literature include independent correlation coefficient, multiple linear regression, logistic regression and canonical correlation analysis. The selection of the procedures depends on the nature of the judgment context. Sometimes, some of these procedures are not applicable. For example, it is not appropriate to use multiple linear regression to capture the policy of decision makers if the situation requires categorical judgment (e.g. pass/fail). In this case, it is advisable to use logistic regression or canonical correlation analysis. Our paper length limits thorough discussion of these procedures. Detailed discussions about the usage of these procedures in judgment analysis contexts can be found in [5].

The number of judgment profiles needed to accurately capture policies using regression analysis has been suggested in judgment analysis literature. Cooksey [5] argues that the number of profiles should have a ratio of 10 to 1 with the number of cues if assuming linear relationship. This ratio can get bigger when attempting to model non-linear relationships. Such ratio have been proposed as a guideline that allows a sufficient number of testing data to unfold the targeted relationship with statistically significant acceptance. This guideline applies to our framework to correctly determine the relationship between the number of tasks required to evaluate users decision policies and the number of insights and findings they reached. Researchers must maintain the ratio dynamically during the evaluation process. Multiple new testing tasks must be conducted as long as the user generates new insights or findings. This explains why it is preferable to conduct a longitudinal study to evaluate the reasoning process of the user of visualization and visual analytics systems.

The coefficients we obtain using regression with raw untransformed samples are called *raw weights*. These weights are not suitable to determine the importance of one cue over another as they are affected by the scale of the variables. That is why it is common to standardize the samples (i.e., transform the vectors $\mathbf{Y_S}$, $\mathbf{Y_e}$ and the columns of the table $\boldsymbol{X}$ to $z$-scores) before applying regression analysis. The resulting coefficients of regression with standardized samples are called $\beta eta$ *weights* which can be used to rank cues based on their contribution to the model of user decisions and the criterion. This provides a quantitative replacement to the qualitative method used in [34] to measure the importance of the insights to experiment subjects and their domain importance. Our quantitative method is also less intrusive as it only uses the data collected over multiple circumstances to define the value of insights.

One of the measures that has been widely used to evaluate the support that the tool provides is the insights count [34]. We argue that it is important to use insights count cautiously. More insights do not necessarily indicate better support. We note that it is possible to generate many unimportant or unrelated insights with some tools and fewer insights with high importance and relevance with other tools. The metrics we have proposed enhance the evaluation of the support that different tools provide by capturing the insights utilization and validation along with their counts.

In this section, we discuss how to capture the user's decision policy and the criterion utilization validity. The issue now is how to compare

them in a more rigorous way. This leads us to the next topic which discusses the decomposition of the achievement correlation to study the situation awareness of a user from different angles.

### 3.2.3 Measuring Awareness

As we noted in the previous section, we need a mechanism to compare judgment policy with the criterion validity (i.e. comparing $f$ with $\bar{f}$). One of the earliest and most beneficial proposals in the field of judgment analysis is the development of Lens Model Equation (LME) that was initially proposed by Hursch *et al.* [19] and extended by Tucker [43]. The equation basically decomposes the Pearson correlation coefficient between the judgment samples and the criterion ($\mathbf{r_a}$) into multiple correlation coefficients that include modeled policy and validity. Several versions of LME have been proposed to perform the same decomposition task but with different types of functions. The version we consider is the original version used when modeling $f$ and $\bar{f}$ as an additive linear functions. LME can be shown as the Equation 1.

$$r_a = GR_eR_s + C\sqrt{1-R_e^2}\sqrt{1-R_s^2} \tag{1}$$

To measure the achievement $\mathbf{r_a}$, we consider four more correlations. The first correlation is the correlation between the judgment $\mathbf{Y_s}$ and the predictions that we can obtain from its model $\hat{\mathbf{Y}}_\mathbf{s}$. This correlation is called cognitive control $\mathbf{R_s}$. Similarly, the correlation between the criterion $\mathbf{Y_e}$ and its model $\hat{\mathbf{Y}}_\mathbf{e}$ can be defined as $\mathbf{R_e}$ which is called the ecological predictability. Squaring these two correlations gives us the percentage of variance in the observed samples that can be explained by the models we capture using regression. The third correlation is $\boldsymbol{G}$ which is the correlation between the two models $\hat{\mathbf{Y}}_\mathbf{s}$ and $\hat{\mathbf{Y}}_\mathbf{e}$. $\boldsymbol{G}$ is commonly called linear knowledge because it compares linear estimation of the judgments with a linear estimation of the criterion. In general, the linear models will not perfectly represent the achievement and to find the correlation between $\mathbf{Y_s}$ and $\mathbf{Y_e}$, we must consider the correlation between samples that are not captured in the linear models. This correlation is called unmodeled knowledge and it is given the symbol $\boldsymbol{C}$. Using these four correlations allows a better diagnosis of decision makers' awareness with respect to certain criterion.

Using Equation 1 allows us to study the contribution of the set of cues (i.e. insights and findings) to the level of awareness achieved by the user. A high value for the unmodeled knowledge $\boldsymbol{C}$ implies two possible scenarios. It can either indicate that the user utilizes the reported cues (i.e. findings and insights) correctly in a non-linear fashion or, more importantly to our evaluation goal, uses information other than the reported cues to make the right decisions. Using this observation allows us to evaluate the qualitative methods used to capture users' insights and findings. Good capturing practice should permit discovering all the information a user observes and utilizes when making a decision. Having high $\boldsymbol{C}$ value indicates that the researcher failed to capture all the insights and findings observed and utilized by the user.

#### *3.2.3.1 Skill Score*

Correlation coefficient can be used as a shape similarity metric but is not sensitive to scale and magnitude differences (see Figure 2 in [20]). It is common to rely on distance measures for more accurate similarity comparison. One of these distance measures is the mean square error (***MSE***) which has been used in meteorology to find the skill of forecasters [25] (see Equations 2,3 and 4). The first equation calculates the distance between the forecasts and the observed events which is an accuracy measure. It is possible to define a distance as a reference or a basis to measure the skill of forecasts. Equation 3 uses the variance of the observed events as a reference distance. We then compare the accuracy and the reference to see how skillful the forecaster is. Equation 4 computes what is called the skill score (***SS***). This metric has an upper bound of 1 when the forecaster predicts the observed event perfectly. There is no defined lower bound to the skill score because, in theory, forecasters can make any prediction.

$$MSE_{accuracy} = \frac{1}{n}\sum(Y_i - O_i)^2 \tag{2}$$

$$MSE_{reference} = \frac{1}{n}\sum(\bar{O} - O_i)^2 \tag{3}$$

$$SS = 1 - \frac{MSE_{accuracy}}{MSE_{reference}} \tag{4}$$

A decomposition of ***SS*** proposed by Murphy [25] who suggests using the correlation coefficient to measure it. Murphy's decomposition (Equation 5) uses correlation coefficient (first term) as a tool to capture the shape similarity between forecasts and observed events and adds two other terms to this to capture the similarity in scale and magnitude. The second term is called regression bias which adds information about scale similarity that is affected by how much a forecaster biases his predictions away from the mean. Mathematically this term captures the effect of having unequal standard deviation for judgment distribution and criterion distribution. Miller *et al.* [23] conducted a study to evaluate how a visualization system can aid humans in reducing this bias as mentioned in the related work section. The last term captures the difference in magnitude that might occur due to over or under estimation.

$$SS = r_{YO}^2 - \left[r_{YO} - \left(\frac{S_Y}{S_O}\right)\right]^2 - \left[\frac{(\bar{Y}-\bar{O})}{S_O}\right]^2 \tag{5}$$

Murphy's decomposition of the skill score is applied in judgment analysis context by Stewart who who further decomposes it by integrating it with LME [40]. The result of this integration can be shown in Equation 6. In this equation, Stewart observed that it is suitable to think about forecasts as judgments which can be analyzed using the lens model. As we noted in LME section, the achievement correlation $\mathbf{r_a}$ is the correlation between judgments and criterion which play the same role of $\mathbf{r_{YO}}$ in Murphy's decomposition. Using LME, Stewart decomposes $\mathbf{r_{YO}}$ into $\mathbf{GR_{OX}R_{YX}}$ where $\mathbf{R_{YX}}$ is the same as the cognitive control of the judge $\mathbf{R_s}$ and $\mathbf{R_{OX}}$ is the same as the ecological predictability $\mathbf{R_e}$.

$$SS = \left(GR_{OX}R_{YX}\right)^2 - \left[r_{YO} - \left(\frac{S_Y}{S_O}\right)\right]^2 - \left[\frac{(\bar{Y}-\bar{O})}{S_O}\right]^2 \tag{6}$$

Equation 6 allows us to evaluate the awareness of VADS users from multiple dimensions. The first term in the equation evaluates user awareness of the trend that the situation follows with respect to different cues' values. The second dimension evaluates the user awareness of the deviation of the criterion. Miller *et al.* [23] define this awareness as the ability to balance between base-rate and case-specific information which controls the deviation from the mean. The last dimension evaluates the awareness of criterion base rate. Increasing the knowledge (awareness) about this information reduces the error that a user can make due to over or under estimation.

#### *3.2.3.2 Uncertainty Impact On User Decisions And Awareness*

Decision making is "the process of sufficiently reducing uncertainty and doubt to allow the reasonable choice of a course of action" [21] [18]. Uncertainty exists as a result of unawareness (or limited knowledge) about a situation. Therefore, we can think of decision making as the process of improving situation awareness.

The uncertainty can be produced and propagated in the visual analytics process and eventually affects user awareness and decisions. Sacha *et al.* [32] used the knowledge generation model to describe the role of uncertainty in the analysis process and the benefit of increasing user awareness about its existence. It is useful to build on their study and show how uncertainty affects user decision policy.

$$SS = \left(GR_{OT}V_{TX}R_{YU}V_{UX}\right)^2 - \left[r_{YO} - \left(\frac{S_Y}{S_O}\right)\right]^2 - \left[\frac{(\bar{Y}-\bar{O})}{S_O}\right]^2 \tag{7}$$

To model the effect of uncertainty on decision policy, we rely on the formulation proposed by Stewart and Lusk [41]. Their work expands

Equation 6 and introduces a version of lens model called the Expanded Lens Model (ELM). The main advantage of this expanded version is in acknowledging the impact of processing and acquiring the information on the awareness level. On one hand, the fidelity of the information system that collects and prepares the information for the decision maker can control delivering true facts accurately or with distortion. Moreover, the accuracy of delivering true facts in the environment to user's cognitive system is affected by the reliability of the information acquisition which represents the effectiveness of a user's perception. To mathematically model this expanded version, Stewart and Lusk expand Equation 6 by adding new parameters to capture the fidelity of the information system $\mathbf{V_{TX}}$ and the reliability of information acquisition $\mathbf{V_{UX}}$. The expanded version is shown in Equation 7.

To conceptualize the effect of uncertainty on the decision policy and the awareness of the user of VADS system, we combine Sacha *et al.* work, with Expanded Lens Model (ELM). Our model can be shown in Figure 3.

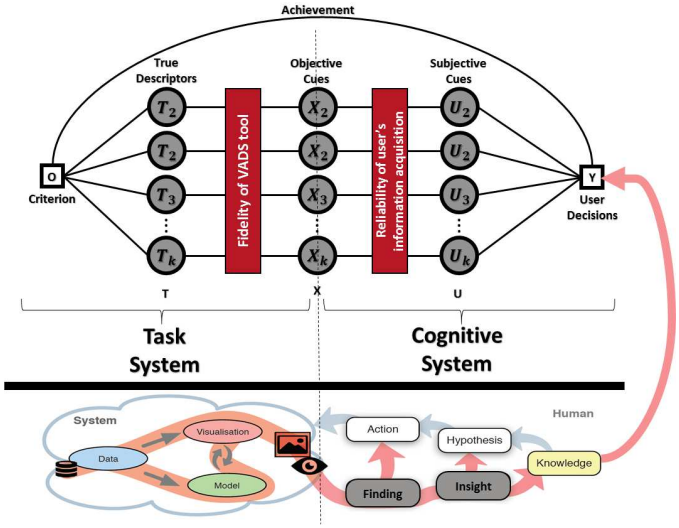


Fig. 3: The effect of uncertainty in VADS system on the information delivery. Above the horizontal line is the ELM [41] and below it is the uncertainity modeling in knowledge generation model [32]. Three levels of cues appear in ELM: True descriptors or facts are the pieces of information as they appear in the environment (the real situation). Objective cues are the same information as the true descriptors with distortion occurring as a result of uncertainty produced or propagated while preparing and processing the information by VADS system. The subjective cues are the same as objective cues with distortion occurring as a result of the unreliability of user perception which affects information acquisition.

The ELM decomposes the proximal part of the environment into three parts as described in the figure. The goal of ELM is to capture the reliability of processing the information in the proximal part and the effect of reliably acquiring this information by the user. Unreliable systems can distort the information and affect decision making. For example, if jigsaw shows that there is a connection between entities X and Y, a user might generate an insight that there is a connection between X and Y. Let's say that this connection has been shown because of an error in defining entities and the real information tells us that there is no such connection. The user perceived distorted information and builds her judgment accordingly. The same issue can happen as a result of the unreliability of the user's information acquisition. The jigsaw user may mistakenly observe from the graph view that there is a connection between A and B because of visual clutter, while the actual situation does not have such connection. Because of this limitation in user vision, the value of the information can become distorted.

Both reliability metrics in ELM are affected by visual analytics systems. The fidelity (reliability) of an information system is affected by the uncertainty produced from the machine side as a result of unreliable data collecting, storing, processing, modeling and visualizing. Moreover, the visualization design can also affect the reliability of acquiring information by the user. For example, the usage of color hue channel has been shown to be more effective in communicating category identification than shape channel as shown in [24]. ELM can conceptually describe the effects of both reliability metrics on the certainty of the information that can be derived from the environment and affect users' decision policies.

The uncertain environments add two roles of VADS systems besides the three already discussed. VADS system should reduce the effect of uncertainty to deliver the information with minimum lost. This can be done by increasing the fidelity of the system and effectively selecting visualization design. The fidelity can be increased in multiple ways such as visualizing the uncertainty effect on communicated information as done by Correa *et al.* [7]. On the other hand, selecting effective visualization can improve the user ability to acquire information.

## 4 Additional Evaluation Metrics

Judgment analysis studies usually are conducted based on one of four possible designs: single system, double-system, triple system and n-system designs [5]. Each of these designs is used to study specific topics. In this section, we use some of these designs to propose evaluation applications of VADS system.

Single system design is a study design which ignores the criterion part in the analysis. It can be used when the goal is to capture user policy only without comparing that policy against any criterion. This is useful in some cases such as evaluating the significance of each cue to the user decision. A common practice in the single system design is aggregating users policies to generalize decision behaviors. This allows studying which insights are likely found in a specific context using a particular VADS system and what common utilization behaviors are.

Double system design is the traditional judgment analysis study design that reflects Brunswik's vision. It includes both the cognitive system to detect a user policy as well as the task system to include the context in the analysis represented by the criterion. We use this design to explain our framework and different concepts throughout the paper. In this section, we are interested in the studies that use the design called multiple cue probability learning 'MCPL'. The goal of this type of study is to evaluate the learning rate of a decision maker by dividing judgment profiles into trials and informing that decision maker about her performance after each trial. Performance feedback is used to teach the decision maker about the difference between her and the criterion which allows us to study the changes in the judgment behavior. We can use the same method to study how a VADS system increases the learning rate by reducing the number of trials needed to achieve certain levels of awareness. We also note that it is possible to use another decision maker policy as the criterion to study how VADS system allows the user to follow the policy of another one.

Previous design can be used to study one sided learning among users. However, a better design to study how multiple users learn from/about each other policies is the triple system or N-system design. These two designs differ in the number of cognitive systems involved in the study. Triple system design studies two cognitive systems with the task system where N-system design studies more than two cognitive systems.

Using triple system design to objectively evaluate collaboration in VADS systems can be an appropriate option. We rely on two types of triple system studies: InterPersonal Conflict studies (IPC) and InterPersonal Learning studies (IPL). The IPC studies includes two stages to study the effect of working with a teammate to change the decision policy. The first stage captures users' policies when they work separately (i.e. $\mathbf{Y_s^1}$ and $\mathbf{Y_s^2}$). The second stage captures the users' joint policy by allowing them to communicate and discuss information and agree on joint decisions ($\mathbf{Y_s^{joint}}$) for the same profiles judged in stage one. This allows studying how the collaboration affect individuals policy favorably or adversely. The IPL design uses the same methodology but with an extra stage which aims to test the knowledge of each decision maker about the policy of his partner. This is done by asking each user to predict his partner's decision after completing stage two. Following

our framework and using IPC and IPL designs allows us to evaluate collaboration in VADS system objectively.

One minor addition to constructing the set of cues needs to be discussed when using IPC and IPL with our framework. As in the single cognitive system case, we propose that the cues are the findings and insights. However, when we capture the policy of the joint decision ($\mathbf{Y_s^{joint}}$) in a multiple cognitive system case, we need to describe it in terms of the complete set of findings and insights that the group reached, i.e., the union of all the findings and insights that each member finds and generates. Using a IPL or IPC design study will allow us to then study how members share findings and insights that influence others in changing their decision policies.

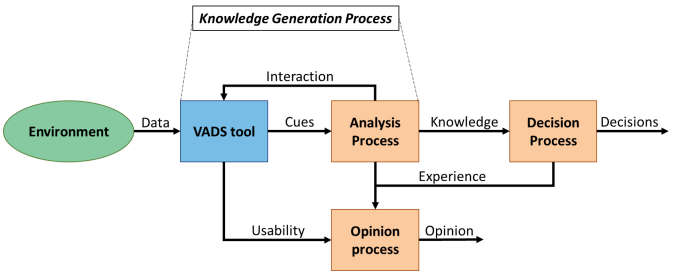


Fig. 4: A model of a VADS system to illustrate the loci of evaluation.

## 5 Comparison With Other Evaluation Frameworks

This section provide a brif comparison between our work and other evaluation frameworks. We focus on the frameworks that are concerned with evaluating visual data analysis and reasoning and user performance (Studies that can be categorized under the second or the fifth scenarios according to Lam *et al.* taxonomy [22]). For comparison, we include MILC [38], traditional insight based evaluation [34], insights and interaction patterns studies such as [29] [16], and performance based evaluation studies such as [46]. The complete comparison is available in Table 1. The table organizes the frameworks on a continuum from highly qualitative to highly quantitative. In this section we explain some of the criteria and terminologies we use in that table.

The intention of evaluation studies can be classified into idiographic and nomothetic. Performance based evaluation studies follow the nomothetic idea. The goal of these evaluation studies is to find the impact of some design choices on the performance of users in general. This explains why we use statistical inferential experiments to prove the study results. On the contrary, when we evaluate our domain expert customer to understand her reasoning process and how to support it as in MILC studies [38], we are following the idiographic idea. The goal in such studies is not to understand the reasoning process of all users in general but rather to capture what affects our customer in particular. If we were to conduct the same study with another customer, our intention would be to find the unique properties of the new customer that are different from what we know about our previous one.

One of the unusual classifications of the approaches that is followed by the frameworks can be observed in our comparison table. The approaches are commonly classified as either qualitative, quantitative, or mixed method approaches. However, we use the term semi-quantitative to describe traditional insight based evaluation studies. Even though, that framework commonly uses insights count as a quantitative measure, it rarely applies any quantitative analysis.

To illustrate the difference between the loci of evaluation for evaluation studies and frameworks, we propose a model for VADS systems. The model treats the components in the system as processes that have inputs, outputs and functions (see Figure 4). The model starts with an environment task system that contains the actual situation. Data is collected from the environment and processed using VADS tools to communicate the situation to the user. The orange components are the user's internal processes. The analytics process is the first process on the human side. This process interacts with the visual analytics tools to find cues that increase the awareness about the environment. The analysis process, along with the VADS tool process, is represented in more detail in the knowledge generation model proposed by Sacha *et al.* [33]. The resulting knowledge from the analysis process is fed into the decision process which utilizes the generated knowledge to make a decision. The opinion of the user about a system is affected by the usability and utility of that system. Utility can be reflected by how good or bad the analysis and decision process are. The opinion of the user can be used to evaluate the tool in some qualitative evaluation methods (e.g. interviews). Domain knowledge affects the functionality of all three orange processes.

One might find our model contradicting with previous studies that suggest a nonlinear or non-sequential relationship between analysis and decision making processes. To resolve such confusion, we need to explain what the decisions in our model are. We think about a decision for a given task as the end choice that is selected by the user from the range of possible answers to that task at the end of the analysis. It should not be confused with temporary decisions that might be made by a user of simulation-like tools to evaluate the consequences of different decisions (tools such as [2]). The later decisions are nothing but interactions that are tested to find more cues and evidence from the supporting tools until a certain level of confidence is established to make the actual decision.

Studies that target evaluating the support of visual data analysis usually focus on understanding the analysis process. Performance based evaluation, on the other hand, focuses on studying the machine component (e.g. VADS tools) objectively. It is common to rely on the outcome of one process to evaluate another one. For example, to evaluate the performance of a user who uses VADS tools, it is possible to measure the accuracy of her using the output of the decision process. Note, however, that the goal of a performance study is not to study the decision process. Similar observation can be seen in MILC which partly relies on expert user opinion process to understand the analysis process.

When we describe the performance based studies procedure in the table, we only present one of the study designs (i.e. the repeated-measures design) to shorten the discussion. Other study designs can change the way the subjects are organized in the experiment which will slightly change performance based section in the table.

The last three criteria in Table 1 are taken from Carpendale's study [4]. Generalizability is defined as the extent to which the result of a study can be seen if the experiment is to be re conducted with new subjects. Realism is defined as the extent to which the framework can be applied in a realistic situation (i.e. real world problem with real users). Finally, precision can be defined as the extent to which an experiment outcome does not change if re conducted with the same setup and subjects. Highly qualitative frameworks have high degree of realism but low degree of generalizability and precision. On the other hand, Highly quantitative frameworks have low degree of realism but high degree of generalizability and precision.

## 6 Challenges And Conclusion

Now, we address some limitations that have not been discussed in this work and can be addressed in future works.

Applying our framework to evaluate decision makers who use visual analytics tools can be costly in terms of time and effort. One challenge is that the framework needs extensive testing data to be applied which increases the time of evaluation experiments and the resources needed to conduct them. The number of judgment profiles must be large enough and should show variability in the environment components to accurately study user decision process idiographically as suggested by Brunswik. We share Brunswik's vision of studying human behavior idiographically, so we consider the practice of extensively testing humans necessary. However, one can make a valid argument by pointing out that large testing data are not always available especially when analyzing a singular non-repeated event.

In this paper, we use one function type (i.e. additive linear) to model the policy and the validity. However, selecting other function types and modifying the proposed formula might be more accurate in some cases. We also model user decision policy without considering the change in

Table 1: Comparison between proposed and common evaluation frameworks

| Framework / Criteria | MILC [38] | Traditional Insight Based Evaluation [34] | Insights Interaction studies [29] [16] | Our framework | Performance Based Evaluation [46] |
|---|---|---|---|---|---|
| **Goal** | Study how integrated visualization system as a whole support analytic process | | Study how users reach insights | Study user utilizations of insights when making decisions about a particular situation. | Study how a particular design choice affects user performance over another |
| **Intention** | Idiographic | Nomothetic | Nomothetic | Idiographic first then Nomothetic | Nomothetic |
| **Approach** | Qualitative | Qualitative and semi-Quantitative | Mixed | Mixed | Quantitative |
| **Loci of evaluation** | Analysis Process | Analysis Process | Analysis Process | Decision process | Machine side tools |
| **Assessment Metrics** | • Performance<br>• Interface efficacy and utility | • Insights importance and value (Qualitative)<br>• Insight count (Quantitative) | • Correlation between insights and interaction pattern<br>• Frequency of moving from interaction pattern to generated insights | • Situation awareness<br>• Learning rate<br>• Collaboration (conflict, learning and joint decision) | • Task completion time<br>• Task accuracy |
| **Number of tools or design choices to evaluate** | **1** tool or ***m*** tools | **1** tool or ***m*** tools | **1** tool or ***m*** tools | **1** tool or ***m*** tools | ***m*** tools |
| **Number of required subjects** | **1** user or ***n*** users | ***n*** users | ***n*** users | **1** user or ***n*** users | ***n*** users |
| **Number of required testing data per subject** | Qualitative data (as many as possible) | ***k*** insights and **1** insights count | ***p*** interaction patterns and ***k*** insights | ***N* instances of:** decision and ***k*** insights | ***m* instances of:** completion time or error rate |
| **Procedure overview** | The researcher get involved and in a real problem with real users and qualitatively extract as much information as possible about the targeted metrics. | The researcher qualitatively collects ***k*** insights for each subjects and qualitatively measures their importance and value to users and to the domain. Total insights count for all subjects are usually calculated to indicate the amount of support that a specific tool provides. | The researcher qualitatively collects ***k*** insights for each subjects and defines ***p*** interaction patterns and then codes this information. The researcher then quantitatively analyzes the relationship between insights and interaction patterns to capture which patterns leads to which insights. | The researcher qualitatively collects ***k*** insights and a decision corresponds to the insights over ***N*** instances to capture situation variability. The researcher then quantitatively analyzes the relationship between insights and decisions to capture the subjects decision policy and awareness. | In repeated-measures design, the researcher collect ***m*** instances of ***n*** completion time or error rate vectors to represent users performance with respect to ***m*** different tools or design choices. Inferential statistics then applied to generalize the findings. |
| **Generalizability** | Limited | Limited | Medium | Medium (with j-users) | broad |
| **Realism** | Highly applicable | Highly applicable | Highly applicable | Medium | Limited to lab setting |
| **Precision** | Low | Low | Medium | High | High |

that policy with respect to different contexts in the situation. Studies such as [26] show that it is possible for the policy to change from one form to another according to different contexts. One of the future works we consider is to study how to enhance our policy modeling by including context parameters as well as other function forms.

The framework proposed in this paper bridges the gap between the field of judgment analysis and the evaluation of VADS systems. We propose using judgment analysis theory to capture the decision policies of users and evaluate the correctness of these policies. It applies to insights based evaluation studies to quantify some of the metrics that have been measured qualitatively. Our framework also introduces some new metrics such as achievement which can be considered a metric to evaluate situation awareness. The description of user decision policy can provide a better diagnosis of a user's awareness level which is an advantage that can complement previously used methods such as SAGAT.

The main contribution of our work can be shown as answers to the following questions. The first question is how a user utilizes the knowledge she generates from interacting with a VADS system when making decisions. The second question we answer using the framework is how to quantify the amount of support that a tool provides to improve the awareness of a decision maker and how to quantify user awareness level itself. The quantitative metrics we have proposed can also help to answer how to compare different solutions and design choices in terms of the amount of support they provides to improve awareness. Finally, we answer the question of why a user achieves a higher awareness level when using one VADS tool instead of another or why users achieve different awareness levels when they use the same VADS tool.

## ACKNOWLEDGMENTS

The authors wish to thank Eileen Arthur for her proofreading.

## REFERENCES


[1] O. Adagha, R. M. Levy, and S. Carpendale. Towards a product design assessment of visual analytics in decision support applications: a systematic review. *Journal of Intelligent Manufacturing*, pp. 1–11, 2015.

[2] S. Afzal, R. Maciejewski, and D. S. Ebert. Visual analytics decision support environment for epidemic modeling and response evaluation. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 191–200. IEEE, 2011.

[3] E. Brunswik. *Perception and the representative design of psychological experiments*. Univ of California Press, 1956.

[4] S. Carpendale. Evaluating information visualizations. In *Information Visualization*, pp. 19–45. Springer, 2008.

[5] R. W. Cooksey. *Judgment analysis: Theory, methods, and applications*. Academic Press, 1996.

[6] J. M. Corbin and A. L. Strauss. *Basics of qualitative research: techniques and procedures for developing grounded theory*. SAGE, 2015.

[7] C. D. Correa, Y.-H. Chan, and K.-L. Ma. A framework for uncertainty-aware visual analytics. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pp. 51–58. IEEE, 2009.

[8] J. W. Creswell. *Qualitative inquiry and research design: Choosing among five approaches*. Sage, 2013.

[9] J. W. Creswell. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications, 2013.

[10] M. R. Endsley. Situation awareness global assessment technique (sagat). In *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National*, pp. 789–795. IEEE, 1988.

[11] O. Gelo, D. Braakmann, and G. Benetka. Quantitative and qualitative research: Beyond the debate. *Integrative psychological and behavioral science*, 42(3):266–290, 2008.

[12] C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. T. Stasko. Jigsaw meets blue iguanodon-the vast 2007 contest. *IEEE VAST*, 7:235–236, 2007.

[13] C. Görg, Z. Liu, N. Parekh, K. Singhal, and J. T. Stasko. Visual analytics with jigsaw. In *IEEE VAST*, pp. 201–202, 2007.

[14] S. Graham and V. S. Folkes. *Attribution theory: Applications to achievement, mental health, and interpersonal conflict*. Psychology Press, 2014.

[15] G. Grinstein, C. Plaisant, S. Laskowski, T. O'Connell, J. Scholtz, and M. Whiting. Vast 2007 contest-blue iguanodon. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pp. 231–232. IEEE, 2007.

[16] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics*, 22(1):51–60, 2016.

[17] M. A. Hardy. *Regression with dummy variables*. Number 91-93. Sage, 1993.

[18] R. Harris. Introduction to decision making. *Home page: http://www.vanguard. edu/rharris/crebook5. htm.[Visited 14 October 2000]*, 1998.

[19] C. J. Hursch, K. R. Hammond, and J. L. Hursch. Some methodological considerations in multiple-cue probability studies. *Psychological review*, 71(1):42, 1964.

[20] A. Kirlik and R. Strauss. Situation awareness as judgment i: Statistical modeling and quantitative measurement. *International Journal of Industrial Ergonomics*, 36(5):463–474, 2006.

[21] J. Kohlhammer, T. May, and M. Hoffmann. Visual analytics for the strategic decision making process. In *GeoSpatial Visual Analytics*, pp. 299–310. Springer, 2009.

[22] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[23] S. Miller, A. Kirlik, A. Kosorukoff, and J. Tsai. Supporting joint human-computer judgment under uncertainty. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 52, pp. 408–412. SAGE Publications, 2008.

[24] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.

[25] A. H. Murphy. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly weather review*, 116(12):2417–2424, 1988.

[26] B. Newsom, R. Mittu, M. A. Livingston, S. Russell, J. W. Decker, E. Leadbetter, I. S. Moskowitz, A. Gilliam, C. Sibley, J. Coyne, et al. Modeling user behaviors to enable context-aware proactive decision support. In *Context-Enhanced Information Fusion*, pp. 231–267. Springer, 2016.

[27] S. Pajer, M. Streit, T. Torsney-Weir, F. Spechtenhauser, T. Möller, and H. Piringer. Weightlifter: Visual weight space exploration for multi-criteria decision making. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):611–620, 2017.

[28] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, vol. 5, pp. 2–4, 2005.

[29] K. Reda, A. E. Johnson, J. Leigh, and M. E. Papka. Evaluating user behavior and strategy during visual exploration. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization*, pp. 41–45. ACM, 2014.

[30] O. C. Robinson. The idiographic/nomothetic dichotomy: Tracing historical origins of contemporary confusions. *History and Philosophy of Psychology*, 13(2):32–39, 2011.

[31] T. L. Saaty. Decision making with the analytic hierarchy process. *International journal of services sciences*, 1(1):83–98, 2008.

[32] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):240–249, 2016.

[33] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, D. Keim, et al. Knowledge generation model for visual analytics. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1604–1613, 2014.

[34] P. Saraiya, C. North, and K. Duca. An evaluation of microarray visualization tools for biological insight. In *Information Visualization, 2004. INFOVIS 2004. IEEE Symposium On*, pp. 1–8. IEEE, 2004.

[35] J. Scholtz. Beyond usability: Evaluation aspects of visual analytic environments. In *Visual Analytics Science and Technology, 2006 IEEE Symposium On*, pp. 145–150. IEEE, 2006.

[36] J. Scholtz, B. Antonishek, and J. Young. Evaluation of a human-robot interface: Development of a situational awareness methodology. In *HICSS*, vol. 4, pp. 50130–3. Citeseer, 2004.

[37] J. C. Scholtz, B. Antonishek, and J. D. Young. Implementation of a situation awareness assessment tool for evaluation of human-robot interfaces. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 35(4):450–459, 2005.

[38] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pp. 1–7. ACM, 2006.

[39] S. S. Stevens. On the theory of scales of measurement, 1946.

[40] T. R. Stewart. A decomposition of the correlation coefficient and its use in analyzing forecasting skill. *Weather and forecasting*, 5(4):661–666, 1990.

[41] T. R. Stewart and C. M. Lusk. Seven components of judgmental forecasting skill: Implications for research and the improvement of forecasts. *Journal of Forecasting*, 13(7):579–599, 1994.

[42] R. Strauss and A. Kirlik. Situation awareness as judgment ii: Experimental demonstration. *International Journal of Industrial Ergonomics*, 36(5):475–484, 2006.

[43] L. R. Tucker. A suggested alternative formulation in the developments by hursch, hammond, and hursch, and by hammond, hursch, and todd. *Psychological review*, 71(6):528, 1964.

[44] M. A. Whiting, J. Haack, and C. Varley. Creating realistic, scenario-based synthetic data for test and evaluation of information analytics software. In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, p. 8. ACM, 2008.

[45] W. Windelband and G. Oakes. History and natural science. *History and theory*, 19(2):165–168, 1980.

[46] Y. Wu, N. Cao, D. Archambault, Q. Shen, H. Qu, and W. Cui. Evaluation of graph sampling: A visualization perspective. *IEEE Transactions on Visualization & Computer Graphics*, (1):401–410, 2017.

[47] E. Zgraggen, A. Galakatos, A. Crotty, J.-D. Fekete, and T. Kraska. How progressive visualizations affect exploratory analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2016.