

# Designing the Evaluation of Operator-Enabled Interactive Data Exploration in VALIDE

Yogendra Patil  
CNRS, Université Grenoble Alpes  
yogendra.patil@univ-grenoble-  
alpes.fr

Sihem Amer-Yahia  
CNRS, Université Grenoble Alpes  
sihem.amer-yahia@univ-grenoble-  
alpes.fr

Srividya Subramanian  
Max-Planck-Institut für  
extraterrestrische  
sri@mpe.mpg.de

## ABSTRACT

Interactive Data Exploration (IDE) systems are technologies that facilitate the understanding of large datasets by providing high level easy-to-use operators. Compared to traditional querying systems, where users have to express each query, IDE systems allow users to perform expressive data exploration following the *click-select-execute* paradigm. Today, there exists no full-fledged evaluation framework for operator-enabled IDE. Most previous works are based on either logging user actions implicitly to compute quantitative metrics or running user studies to collect explicit feedback. Hence, there is a pressing need to articulate an evaluation framework that *collects and compares quantitative human feedback* along with system and data-centric evaluations. In this paper, we develop VALIDE, a preliminary design of a unified framework consisting of a methodology and metrics for IDE systems. VALIDE combines research from database benchmarking and human-computer interaction and will be demonstrated with a real IDE system.

## 1 INTRODUCTION

Recent decades have seen a tremendous rise in the availability of very large datasets ranging from healthcare to sports and social media. This has been accompanied by a rise in data exploration stakeholders with varying expertise in computer science. The field of astrophysics research is no exception. The Sloan Digital Sky Survey (SDSS) is an example of astronomical database commonly used in the astrophysics community explored using SQL [8]. SQL-based IDE systems require users to spend enormous time in training, formulating and refining queries, and utilizing different means for visualizing data samples. To address that, new operator-enabled IDE systems have been developed that do not require users to be familiar with SQL [6, 11, 15, 20, 27]. IDE operators are important tools that provide different data access modalities such as filtering data, finding subsets and supersets, and looking for similar and dissimilar items. However, there exist no framework that performs an end-to-end evaluation of such operator-enabled IDE systems. In this paper, we develop VALIDE (eVALuation of IDE systems), a general-purpose evaluation framework for IDE and illustrate its

applicability using one such operator-enabled system, DORA THE EXPLORER [27]. However, evaluating an IDE system must address three key dimensions: data, system, and human. The data dimension assesses the ability of the system to allow users to find dispersed data in a very large dataset. The system dimension reflects performance such as query execution time. The human dimension captures user interactions and perception with the system during the exploration session.

VALIDE is a unified framework that investigates three key dimensions of an IDE system – data, system, and human. To explore each dimension, we design dedicated metrics. A ‘data’ metric quantifies recovered ‘data’ during an exploration session via measuring the closeness of user discovered and system recommended datasets to some ground truth. A ‘system’ metric quantifies the performance of the front-end and back-end of an IDE system. A ‘human’ metric characterizes the human interactions with the IDE system quantitatively. Quantitative human metrics generally refer to quantifiable human actions during data exploration. We further classify human metrics as – Human System Interactions or **HSI** (e.g. the time spent by a user to complete a data exploration task) and Human Self Evaluations or **HSE** (e.g. the feeling of accomplishment reported by the user on 5-point Likert scale). Compared to other work, our study collects both types of human metrics and computes them from user sessions by performing factorial design [17, 28]. The goals of VALIDE can be summarized as follows:

- (1) *Goal# 1:* Analyze data, system & human metrics in an operator-enabled IDE system using robust statistical techniques for human data collection,
- (2) *Goal# 2:* Design a generic framework for IDE evaluation with high level operators (or querying methods),
- (3) *Goal# 3:* Develop an end-to-end methodology that consists of training & testing subjects, collecting both human quantitative metrics (**HSI** and **HSE**), and cross-checking them,
- (4) *Goal# 4:* Demonstrate human-centric evaluation by applying it on an operator-enabled IDE system.

### 1.1 Challenges and Contributions

VALIDE distinguishes itself by capturing essential aspects of user perception of an IDE system (e.g. Mental Demand<sup>1</sup>), which are not captured with other metrics [9, 28, 32]. Consider a data exploration session, where an astrophysicist is interested in finding a specific set of galaxies. From the session, the derived data, system and **HSI** metrics revealed that the astrophysicist was able to recover all data subsets of interest, with minimal query execution delays, and a few

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

HILDA '22, June 12, 2022, Philadelphia, PA, USA

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9442-0/22/06...\$15.00  
<https://doi.org/10.1145/3546930.3547509>

<sup>1</sup>refers to the amount of work where the execution of a specific task requires that subject perform mental processes, such as thinking, deciding, calculating, remembering, analyzing, searching, observing, to mention few [31].

exploration steps. But, before launching the IDE task, the astrophysicist may have to put a lot of effort getting familiarized with the IDE system. Hence, the results from data, system, **HSI** metrics may favor the use of new functionalities (e.g., expressive operators), but the astrophysicist's perception may not be positive. Also, as compared to previous studies which have fragmented usage of data, system, human - **HSI**, **HSE** metrics, VALIDE tries to create a unified generic framework with humans in the loop. Therefore, *the first challenge* lies in capturing user perception of the IDE system. Previous researchers addressed the evaluation of IDE systems e.g. cross-filtering using slider, by calculating data, system, and **HSI** metrics [28]. They evaluate IDE based on workloads and interactions, but fail to capture important aspects such as mental demand and perceived controllability. *The second challenge* is how to make sound observations from surveys, given that human behavior is unpredictable and simply reporting averages is bound to produce errors and outliers. We propose to make use of factorial design, a method that is widely adopted in studies involving human participants [9]. *The third challenge* lies in verifying feedback, as human subjects are prone to bias, e.g. when asked about their quality of work [16]. To address that, we propose to cross-check self-reported feedback (**HSE** metrics) with human exploration activity (**HSI** metrics).

## 2 RELATED WORK

Our research work builds upon existing work related to evaluation of IDE for large datasets.

**IDE evaluation with system metrics:** A commonly observed evaluation method for an IDE system involves applying Transaction Processing Performance Council (TPC) suites [2]. Depending upon the specific workload type and application requirements, a type of TPC suite is selected. Applied metrics primarily focus on capturing response time and quality of results [18]. Eichmann et al [10] argue that TPC is useful for data exploration systems that produce static workloads and propose a method to evaluate dynamic workloads. Another study [7] improves upon this by generating dynamic workloads from simulated agents (developed from pre-recorded user exploration sessions) and addition of new metrics that capture response time and quality of results. The goal of these studies is to evaluate the query processing engines, but they do not include understanding user perception in IDE.

**IDE evaluation with system & human metrics:** Rahman et al. [28] categorizes evaluation metrics used for an IDE system into system and human metrics. The former mainly focus on capturing response time (e.g., time delays in query scheduling and processing), while the latter focus on quantifying human behavior (e.g. exploration duration) and satisfaction by deploying user surveys. This study demonstrates the application of these metrics to evaluate different IDE querying interfaces by human subjects. A similar study by Jiang et al. [14] evaluates user experience by using unique system metrics that characterize the effect of the workload, created by the interactive system mechanism, on response time. Although both studies [14, 28] outline various methods for evaluating querying interfaces for an IDE system, they calculate simple statistics (e.g. mean, count, etc.) to assess human behavior and derive conclusions. Such methods are prone to errors as human behavior is noisy and needs a more robust statistical modeling methodology. Closer to

**Table 1: The NASA-TLX questionnaire used in our study for understanding user perception in IDE.**

Type	Question
Feeling of Accomplishment (Q1)	How successful were you in accomplishing what you were asked to do?
Effort Required (Q2)	How hard did you have to work to accomplish the task?
Mental Demand (Q3)	How mentally demanding was the task?
Perceived Controllability (Q4)	How discouraged, irritated, stressed, and annoyed were you?
Temporal Demand (Q5)	How hurried or rushed was the pace of the task?

our research is the work performed by Abouzied et al [3] and Liu et al [21]. Abouzied et al [3] introduce an IDE querying tool that simplifies the specification of complex SQL queries by allowing users to directly manipulate the query or apply auto-correction. However, this study performs evaluation using human metrics by applying ANOVA analysis to model human data only and does not include system or data metrics. Liu et al [21] introduce a query interface for non-technical database users. The evaluation process consists of recruiting participants without SQL background. Certain **HSI** metrics were derived along with participant opinions using feedback. Here again, the analysis involved calculating simple statistics (e.g. count) to evaluate human behavior and derive conclusions.

Therefore, to the best of our knowledge, previous works on IDE evaluation mainly focus on either system or human metrics (**HSI**). Out of them, very few implement a robust statistical analysis to derive conclusions [28]. None of these studies develops a generic and unified evaluation framework that employs data, system and human metrics and analyzes them together.

## 3 VALIDE FRAMEWORK

This section summarizes the generic experiment design and methodology of VALIDE (*Goals# 2 and # 3* in Section 1). In Section 5 we show how it is deployed on a specific system.

### 3.1 Design of Use Cases

The first step is to define a training and a testing use case. Training use case allows participants to get familiarize with the IDE system, whereas test use case requires participant to perform an IDE task without external assistance. The train and test use cases must be mutually exclusive cases and commonly observed real-world scenarios designed in collaboration with an expert.

### 3.2 Design of Study

Our study follows a 2x2 factorial design and “between subjects”, so as to mitigate the “learning effect” [16]. Factorial design is a widely adopted experimental design for understanding the effect of a set of independent variables or **factors** on dependent variables. The experimenter first holds other factors at constant level while varying the factors under consideration. Then the experimenter assigns participants to various groups containing factors under consideration at certain values. This allows to model human behavior based on only the factors under consideration and, thus no other factors can interfere with the analysis [17].

### 3.3 Deriving Human Quantitative Metrics

The National Aeronautics and Space Administration-Task Load Index (NASA-TLX) is a widely used tool for understanding human perception related to a given task through a set of carefully designed questions [1]. In this work, we adapt NASA-TLX as shown in Table 1. Each question allows to derive a **HSE** metric on a Likert scale between 1 (unfavorable) and 5 (favorable).

Another way to derive human quantitative metrics is by logging various user interactions, defined as **HSI** metrics in Section 1. This allows to characterize the human behavior quantitatively using the logged data [24, 25, 33, 34]. This, in-turn, allows to quantify user sessions with data exploration systems and verify participants' feedback on NASA-TLX questionnaire. Based on previously conducted research studies, we define various metrics to characterize IDE session [7, 10, 28]. The **HSI** metrics derived by logging user sessions are:

- (1) **Length of Pipeline** defines the total number of operations contributing to the exploration of the target data subset.
- (2) **Total Exploration Duration** expresses the difference between the time at which the user submits the first query and the time at which the user ends the data exploration task.
- (3) **Average Time Per Step** defines the average time taken by the data explorer to submit each query at each step.

### 3.4 Deriving System and Data Metrics

We define the following data and system metrics:

- (1) **Recall Rate**: Proportion of relevant data in the recovered data.
- (2) **Precision Rate**: Proportion of recovered data that are relevant
- (3) **Average Query Delay** represents the average time duration taken between submission and execution of a query.

### 3.5 Analysis of Metrics

In case of **HSE**, since the response is a categorical variable (5 point Likert scale), a Kruskal-Wallis significance test is used to check if there was a difference in sample means of responses across groups. Then the Wilcoxon rank-sum test should be applied to compute the pairwise statistical significance [29]. We set our null hypothesis ' $H_0$ ' as- "For a given NASA-TLX question, there is no significance difference between responses of participants from various groups". In case of data, system, and **HSI**, since the values are continuous variables, a two-way ANOVA significance test is utilized to check if there is a difference in sample means across all groups. Then a Tukey's Honestly Significant Difference (HSD) test is applied to find out the pairwise statistical significance [22]. We set our null hypothesis ' $H_0$ ' as- "For each metric, there is no significant difference in values from various groups".

Now consider that a user provides feedback on Feeling of Accomplishment after completing the given IDE task. To cross-check this metric, we propose to utilize Recall Rate which is derived from the logged session for that particular user. So, if the user provides a Likert score of 5 (most favorable) for Feeling of Accomplishment, then correspondingly the value for Recall Rate should be 100%. Similarly, for user feedback on Effort Required, we propose Length of Pipeline, and for Temporal Demand, we propose Total Exploration

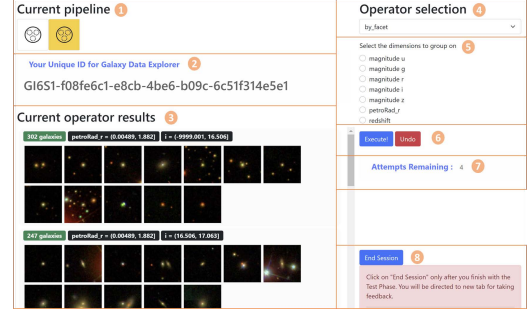


Figure 1: Modified front-end design of DORA THE EXPLORER for experimentation.

Duration for cross-checking (Goal# 3). However, for the metrics Mental Demand and Perceived Controllability, it is not possible to determine their counterparts as they require special intervention like application of special medical devices (e.g. EEG) and cannot be simply cross verified by logging participants' interactions with the IDE system [9, 32].

## 4 VALIDE WITH DORA THE EXPLORER





To achieve Goal# 1 & 4, we apply our framework on an operator-enabled IDE system – DORA THE EXPLORER<sup>2</sup> designed for SDSS data exploration [26, 27]. We represent the dataset as a set of records  $\mathbb{D}$ . Each record describes a galaxy with a set of 7 attributes  $A$ : "u, g, r, i, z" describes the photometric magnitudes and brightness of galaxies in SDSS filters, "petroRad\_r" describes the size of galaxies, and "redshift" characterizes the spectroscopic redshifts and measures the distance of galaxies from the Earth. In general, a galaxy can be distinguished by its color, shape and structure. Figure 1 shows two samples produced with DORA THE EXPLORER (under Module 3 - Current operator results) containing 302 & 247 galaxies described by conjunction of 2 attribute values or simply  $a_m = \{\text{petroRad}_r, i\}$ , where  $0 < m \leq n$ . An *exploration pipeline* is a sequence of operators whose purpose is to recover a desired data subset. In its general form, an operator takes a set of objects  $\mathcal{D} \subseteq \mathbb{D}$  and returns sets of objects that are related to objects in  $\mathcal{D}$ . Table 2 summarizes the set of operators used in DORA THE EXPLORER. Since they are applied to sets, we represent their equivalent definition in the *Region Connection Calculus 8* (RCC8) formalism [19, 30] (second column). Figure 1 presents the modules of DORA THE EXPLORER which we extended for our study.

### 4.1 Use Cases for DORA THE EXPLORER

Our use cases were provided by an astrophysicist (a co-author of this paper), well-versed in SQL querying over SDSS. The training use case consists of - (1) pointed shape, (2) located far away from Earth, and (3) emitting light primarily in (i) visible (ii) near infra-red wavelengths galaxies. To find them, our expert writes a rough SQL query, extracts a first sample dataset and then uses the SkyServer

<sup>2</sup>With the developers' consent, we use DORA THE EXPLORER in our study to demonstrate the application of VALIDE. The code is freely available at <https://github.com/apersonnaz/rl-guided-galaxy-exploration>, and the application at <https://bit.ly/dora-application>

**Table 2: Examples of exploration operators in DORA THE EXPLORER: input in bold lines, output in dashed lines.**

Operator	RCC8 Formalism [30]	Output description
by-facet( $D, A$ )	NTPP1 	returns as many subsets of $D$ as there are combinations of values of attributes in $A$
by-superset( $D, k$ )	NTPP 	returns the $k$ smallest supersets of input set $D$ ( $k$ is application-dependent)
by-distribution( $D$ )	DC 	returns all sets that are distinct from the input set $D$ and whose attribute value distribution is similar to $D$
by-neighbors( $D, a$ )	EC 	returns 2 sets that are distinct from the input set $D$ and that have the previous (smaller) and next (larger) values for attribute $a$

Imaging Query Form [8] to check visually if the sampled dataset contains Pointed Shape galaxies and further refines SQL queries to find as many galaxies of interest as possible.

```
SELECT count(*)
FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
AND p.petroRad_r BETWEEN 1.882 AND 2.759
AND p.redshift BETWEEN 0.201 AND 0.333
AND s.z BETWEEN 16.506 AND 16.753
AND p.r BETWEEN 16.928 AND 17.496
```

**Listing 1: An expert-created SQL query for finding Pointed Shape galaxies**

For the test use case, we consider galaxies which are - (1) edge-on spiral shape, (2) located near-by, and (3) emitting light in (i) visible, (ii) near infra-red range. This identifies Spiral edge-on galaxies whose query is:

```
SELECT count(*)
FROM PhotoObj AS p
JOIN SpecObj AS s ON s.bestobjid = p.objid
AND p.petroRad_r BETWEEN 4.55 AND 258.486
AND p.redshift BETWEEN -0.01 AND 0.0781
AND s.z BETWEEN -9999.001 AND 16.753
AND p.r BETWEEN 16.928 AND 17.496
```

**Listing 2: An expert-created SQL query for finding Spiral Edge-on galaxies**

## 4.2 Study Design for DORA THE EXPLORER

We define two exploration modes shown in Table 3. Traditional Operator represents an exploration mode limited to using by-facet to search for subsets of an input set (akin to drill-down) or using by-superset to search for supersets of an input set (akin to roll-up). These exploration modes are further extended by the All Operators set by adding by-neighbors and by-distribution. The factor **total interactions** reflects the total number of query submissions or query undos that the user is allowed in a session. For **total interactions**, the levels are MIN and MAX. MIN is a lower-bound and represents a well-written expert SQL query for finding the required dataset. MAX is an upper-bound and corresponds to a non-expert written SQL query for finding the required dataset. We set MIN to '7' -

**Table 3: Levels for factors, operators and total interactions.**

Levels	Description
All Operators	A set of all operators as defined in Table 2
Traditional Operators	A set of operators consisting of only by-facet & by-superset
MIN	represents the total clauses & attributes in an expert-created SQL query
MAX	represents an upper bound on a non-expert SQL query attempts.

same as the total number of SQL clauses and attributes required by our expert to write the SQL query 2 in Section 4.1. We set MAX to '21' to reflect the maximum number of attempts in the SDSS logs<sup>3</sup> required for a non-expert to create the same SQL query<sup>4</sup>.

## 4.3 Deriving Human Metrics

The HSI metrics derived by logging user sessions are:

- (1) **Length of Pipeline** defines the total number of operators contributing to the target data subset. Suppose the user performs a total of  $l$  operations to get the desired result, and during this the user performs various undos, then 'Length of Pipeline'  $L_p$  is expressed as:

$$L_p = \sum_{i=1}^l O \begin{cases} O = +1, \text{ user executes an operation} \\ O = -1, \text{ user undos an operation} \end{cases} \quad (1)$$

- (2) **Total Exploration Duration** expresses the difference between the time at which the user submits the first operator  $t_s$  and the time at which the user ends the data exploration task  $t_e$ . Total Exploration Duration ( $T_E$ ) is denoted as:

$$T_E = t_e - t_s \quad (2)$$

- (3) **Average Time Per Step** defines the average time taken by a user to submit an operator at each step. Consider a set  $\mathcal{T}$  containing records of timestamps at which the user submitted each operator, i.e.  $\mathcal{T} = \{t_1, t_2, \dots, t_n\}$  then Average Time Per Step  $T_{ps}$  is given as:

$$T_{ps} = \frac{\sum_{i=1}^{n-1} t_{i+1} - t_i}{n - 1} \quad (3)$$

## 4.4 Deriving System and Data Metrics

We define the following data and system metrics:

- (1) **Recall Rate**: Consider where the user finds a dataset  $\mathcal{D}(a_j)$ , using a set of attributes  $a_j = \{z, \text{petroRad}_r, \text{redshift}, i\}$ , where  $a_j \in A$  and  $0 < j \leq n$ . Now, the gold standard dataset for  $a_j$  is defined by expert SQL query as  $\mathcal{G}(a_j)$ . Hence, the "Recall Rate" or  $R(a_j)$  is expressed as:

$$R(a_j) = \frac{\mathcal{D}(a_j) \cap \mathcal{G}(a_j)}{\mathcal{G}(a_j)} \quad (4)$$

<sup>3</sup>Attempts here signifies re-formulation of query at each step due to typos, missing predicates, and missing or in-accurate operators.

<sup>4</sup>See the query interface and logs produced here <http://skyserver.sdss.org/log/en/traffic/sql.asp>

- (2) **Precision Rate:** Consider where the user finds a dataset  $\mathcal{D}(a_j)$ , using a set of attributes  $a_j = \{z, \text{petroRad}_r, \text{redshift}, i\}$ , where  $a_j \in A$  and  $0 < j \leq n$ . Now, the relevant items in the recovered dataset  $\mathcal{D}(a_j)$ , defined by an expert SQL query, as  $\mathcal{R}(a_j)$ . Hence, the “Precision Rate”<sup>5</sup> or  $P(a_j)$  is expressed as:

$$P(a_j) = \frac{\mathcal{D}(a_j)}{\mathcal{D}(a_j) \cup \mathcal{R}(a_j)} \quad (5)$$

- (3) **Average Query Delay** represents the time duration between submission and execution of a query. Given a user has executed ‘ $n$ ’ queries and  $Q_{di}$  represents query delay for the  $i^{th}$  query,  $Q_d$  is expressed as:

$$Q_d = \frac{Q_{d1} + Q_{d2} + \dots + Q_{dn}}{n} \quad (6)$$

## 5 DEPLOYMENT OF EXPERIMENT

We now describe our experiment design (*Goals# 4* in Section 1). In total 196 participants were recruited on Prolific Academic [23] from various academic background<sup>6</sup>. Subjects recruited were first redirected to “Google Forms” to complete the study consent form. Our study was carried out following the guidelines of the Data Protection Officer at our university and as governed by the European Commission’s standard contractual clauses. During the entire study, there was no direct in-person contact with the participants and their identity was concealed at all times. Due to that, we first conducted a pre-screening attention test. We asked participants to watch a presentation for 10 mins related to background details of data exploration and asked to answer 5 multiple-choice questions related to the presentation. Out of 196, 89 participants demonstrated their complete attention and commitment by answering all questions correctly. Finally, 84 of them consented to continue with the study. Participants were separately rewarded with monetary means at completion of the pre-screening test and the study. Given that the population size (number of stakeholders in galaxy data exploration) is large, and the sample size is 84, the confidence level achieved is 90%, with a margin of error of about 9% [12]. Each participant was assigned to one group (in Table 4) using the proportional stratified sampling method [16]. This assures that each group is assigned with the same representative sample from the total set of participants. Participants assigned to each group were first redirected to “Google Forms” to complete the study consent form and then to follow an online training video so as to train participants to use DORA THE EXPLORER using the training use case. Subjects were trained to use the different operators with rough estimates of values for the attributes as described in Section 4. This process mimics real-world scenarios with astrophysicists. After training, each user was asked to complete the test use case.

## 6 RESULTS AND DISCUSSION

We demonstrate the analysis of data, system, and human metrics. Results for each metric are presented for treatment ‘groups’ described in Table 4.

<sup>5</sup>In case of DORA THE EXPLORER  $\mathcal{D}(a_j) \cup \mathcal{R}(a_j) = \mathcal{D}(a_j)$ , hence Precision Rate is always 1 regardless of participant group, therefore is not considered for statistical analysis purpose.

<sup>6</sup>inline with goals described in [4, 5]

**Table 4: List of groups assigned with different levels of factors used in this study.**

Group	Treatments
Group-1	All Operators with MIN Interactions
Group-2	Traditional Operators with MIN Interactions
Group-3	All Operators with MAX Interactions
Group-4	Traditional Operators with MAX Interactions.

### 6.1 Analysis of HSE Metrics

Figure 2 reports the mean value for participants’ response to the NASA-TLX questionnaire from each group. Results from Kruskal-Wallis significance test indicate that a significant effect was observed only for Q1 - Feeling of Accomplishment with p-value=0.003. In this case, we ran pairwise comparisons using a Wilcoxon rank-sum test with Bonferroni correction of  $\alpha = (0.05 \div \text{total treatments}) = (0.05 \div 4) = \mathbf{0.01}$ . Further, running a pairwise comparison between group responses for Q1 using a Wilcoxon rank-sum test indicated a significant effect between Group-1 (All Operators and MIN interactions) vs Group-2 (Traditional Operators and MIN interactions), Group-1 vs Group-4 (Traditional Operators and MAX interactions), Group-2 vs Group-3 (All Operators and MAX interactions), and Group-3 vs Group-4. This reveals that the significant difference between participants of Group-1 & 3 vs Group-2& 4 mainly appears when they use All Operators. For the remaining HSE metrics, results indicate that participants were able to use DORA THE EXPLORER with – (1) ease to complete the task in the allotted time and training (Temporal Demand), (2) some to regular effort (Effort Required), (3) less stress to some stress (Mental Demand), and (4) little to very little assistance (Perceived Controllability).

### 6.2 Analysis of Data, System, HSI Metrics

From Figure 3 (a) it can be observed that on average, participants of Group-3 have a high Recall Rate (66 %) comparatively to others. To check if there is statistically significant differences in Recall Rate or **dependent variable** due to variation in factors or **independent variable** – operators and total interactions, various statistical models were developed. These models were then tested using - a one way ANOVA (to test individual model containing each factor), a two-way ANOVA (to test a single model containing both factors), a two-way ANOVA with interactions<sup>7</sup> (to test a single model containing both factors and interactions terms). Results indicate a statistically-significant difference in mean values of Recall Rate only when levels for factor ‘operators’ were varied, i.e. for a model with ‘operators’ coded as categorical independent variable, with  $(F(1,80) = 4.145, p < 0.04^*)$ . Further analysis of this model using Akaike information criterion (AIC) revealed an AIC weight of 46% (i.e. it explains 46% of the total variation in the dependent variable - Recall Rate). Finally, it was verified that the model fits the assumption of homoscedasticity using diagnostic plots [13]. We conducted a Tukey’s HSD post-hoc test for pairwise comparisons between the two levels of operators - All Operators and Traditional Operators [22]. The post-hoc test results indicate that there

<sup>7</sup>the term interactions is widely used in ANOVA analysis and should not be confused with our factor ‘total interactions’.



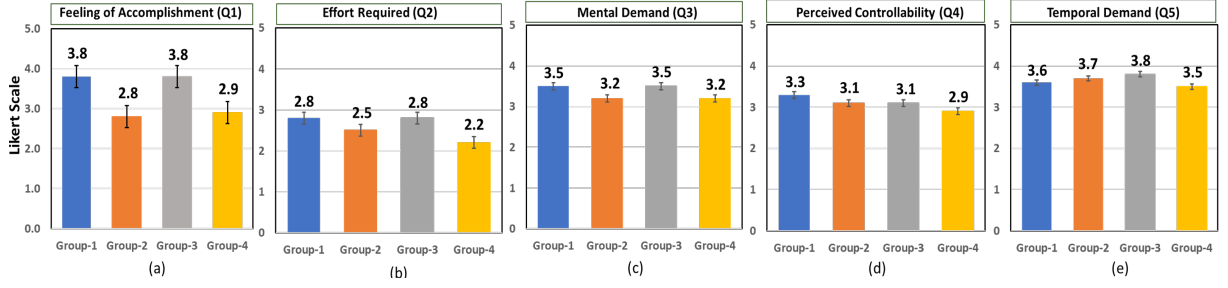


Figure 2: Plots for mean value of user feedback for each NASA-TLX questionnaire on Likert scale (the only significant result is Feeling of Accomplishment (p-value=0.003)).

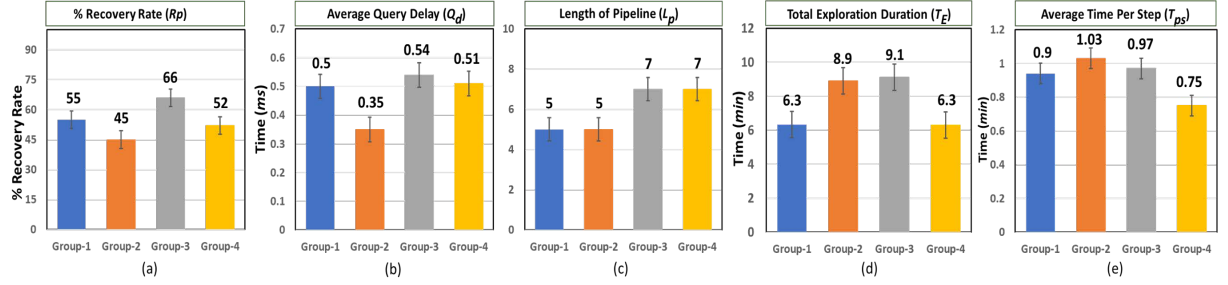


Figure 3: Plots for calculated mean values of quantitative metrics from logs of each participant session.

is statistically-significant differences ( $p < 0.044$ ) between groups with All Operators and Traditional Operators and revealed that All Operators set resulted in an increase of Recall Rate on average than Traditional Operators by 13.64 units. Cross checking this with HSE metric analysis for Q1 - Feeling of Accomplishment does confirm our claim that All Operators does in fact assist in higher recovery of target data.

Similarly, this entire process was repeated for the remaining metrics individually in order to understand the effect of varying the factors or **independent variable** – operators and total interactions on **dependent variables** – Average Query Delay ( $Q_d$ ), Length of Pipeline ( $L_p$ ), Total Exploration Duration ( $T_E$ ), Average Time Per Step ( $T_{ps}$ ) individually. Analysis indicates that there was no significant effect caused by varying the independent variables on any of the dependent variables except for Length of Pipeline. There is a statistically-significant difference in Length of Pipeline only when the factor total interactions were varied, with an AIC weight of 68% and fitting the assumption of homoscedasticity. The Tukey’s HSD post-hoc test indicates statistically significant differences ( $p < 0.0103$ ) and reveals that MAX interactions resulted in an increase of Length of Pipeline on average than MIN interactions by 1.6 units, i.e. nearly 2 extra operators were used by participants that were assigned to groups having total interactions as MAX. This indicates that participants tend to explore further when the total interactions allowed is increased. Now, combining results, although participants perceive Effort Required (as per HSE metric analysis) is similar across different groups, the HSI metric analysis showcases that participants from Group-3 and Group-4 tend to apply more operators (by definition of Length of Pipeline). One reason for the discord between Effort Required and its corresponding quantitative metric

Length of Pipeline is that the difference (2 extra operators) is small thereby incurring a similar perceived effort.

## 7 CONCLUSION AND FUTURE WORK

We developed VALIDE, a preliminary human-centric evaluation framework for operator-enabled IDE. We outlined the general framework for VALIDE and demonstrated its application to DORA THE EXPLORER. VALIDE consisted of metrics and a methodology that enable cross-checking with human feedback. VALIDE allows to understand some inherent limitations in gathering and relying on user feedback only. For example, participants across different groups reported that there were no overall significant differences in “Effort Required”, however cross checking this with “Length of Pipeline” (HSI) metric, there appears to be some discrepancies. Hence, VALIDE allows to point out such cases. Application of factorial design along with statistical tests allowed us to conclude sound results. Our results on DORA THE EXPLORER indicated that participants using All Operators were able to improve the Recall Rate with respect to a ground truth, while an increase in total human-data interactions had no significant effect on Recall Rate. However, the best observed Recall Rate was 66%. This may be because users cannot intervene and modify the semantics of available operators for DORA THE EXPLORER. Also, it may be due to the lack of assistance in the form of (data subset, operator) recommendation at each exploration step. In addition, VALIDE largely focused on manual exploration. Thus, the evaluation using VALIDE opens two new directions: extending DORA THE EXPLORER with user interventions to modify operators and support new visualizations, and going beyond manual exploration by enabling stepwise and end-to-end recommendations in IDE.

## REFERENCES

- [1] 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183.
- [2] 2017. Oracle TPC benchmark. <http://www.tpc.org>.
- [3] Azza Abouzied, Joseph Hellerstein, and Avi Silberschatz. 2012. Dataplay: interactive tweaking and example-driven correction of graphical database queries. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 207–218.
- [4] Sihem Amer-Yahia, Georgia Koutrika, Frederic Bastian, Theofilos Belmpas, Martin Braschler, Ursin Brunner, Diego Calvanese, Maximilian Fabricius, Orest Gkini, Catherine Kosten, Davide Lanti, Antonis Litke, Hendrik Lücke-Tieke, Francesco Alessandro Massucci, Tarcisio Mendes de Farias, Alessandro Mosca, Francesco Multari, Nikolaos Papadakis, Dimitris Papadopoulos, Yogendra Patil, Aurélien Personnaz, Guillem Rull, Ana Sima, Ellery Smith, Dimitrios Skoutas, Srividya Subramanian, Guohui Xiao, and Kurt Stockinger. 2021. INODE: Building an End-to-End Data Exploration System in Practice [Extended Vision].
- [5] Sihem Amer-Yahia, Georgia Koutrika, Martin Braschler, Diego Calvanese, Davide Lanti, Hendrik Lücke-Tieke, Alessandro Mosca, Tarcisio Mendes de Farias, Dimitris Papadopoulos, Yogendra Patil, Guillem Rull, Ellery Smith, Dimitrios Skoutas, Srividya Subramanian, and Kurt Stockinger. 2022. INODE: Building an End-to-End Data Exploration System in Practice. *SIGMOD Rec.* 50, 4 (jan 2022), 23–29. <https://doi.org/10.1145/3516431.3516436>
- [6] Sihem Amer-Yahia, Tova Milo, and Brit Youngmann. 2021. Exploring Ratings in Subjective Databases. In *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, Guoliang Li, Zhanhui Li, Stratos Idreos, and Divesh Srivastava (Eds.). ACM, 62–75.
- [7] Leilani Battle, Philipp Eichmann, Marco Angelini, Tiziana Catarci, Giuseppe Santucci, Yukun Zheng, Carsten Binnig, Jean-Daniel Fekete, and Dominik Moritz. 2020. Database benchmarking for supporting real-time interactive querying of large data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1571–1587.
- [8] Michael R. Blanton and et al. 2017. Sloan Digital Sky Survey IV: Mapping the Milky Way, Nearby Galaxies, and the Distant Universe. 154 (2017).
- [9] Brad Cain. 2007. A Review of the Mental Workload Literature.
- [10] Philipp Eichmann, Emanuel Zraggen, Carsten Binnig, and Tim Kraska. 2020. Idebench: A benchmark for interactive data exploration. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 1555–1569.
- [11] Apostolos Glenis and Georgia Koutrika. 2020. NLonSpark: NL to SQL translation on top of Apache Spark.
- [12] Joe Kotrlik James Bartlett and Chadwick Higgins. 2001. Organizational research: Determining appropriate sample size in survey research appropriate sample size in survey research. *Information technology, learning, and performance journal* 19, 1 (2001), 43.
- [13] Carlos M Jarque and Anil K Bera. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics letters* 6, 3 (1980), 255–259.
- [14] Lilong Jiang, Protiva Rahman, and Arnab Nandi. 2018. Evaluating Interactive Data Systems: Workloads, Metrics, and Guidelines. In *SIGMOD*. 1637–1644.
- [15] Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. 2017. Interactive data exploration with smart drill-down. *IEEE TKDE* (2017).
- [16] Harry Hochheiser Jonathan Lazar, Jinjuan Feng. 2017. *Research Methods in Human Computer Interaction (Second Edition)*. Morgan Kaufmann, Boston.
- [17] S Prion K H Adamson. 2020. Two-by-Two Factorial Design. *Clinical simulation in nursing* 49 (2020), 90–91.
- [18] Samy Kabangu. 2009. Benchmarking Databases. (2009).
- [19] Sanjiang Li and Mingsheng Ying. 2003. Region connection calculus: Its models and composition table. *Artificial Intelligence* 145, 1-2 (2003), 121–146.
- [20] Matteo Lissandrini, Davide Mottin, Themis Palpanas, and Yannis Velegrakis. 2020. Graph-Query Suggestions for Knowledge Graph Exploration. In *WWW*. 2549–2555.
- [21] Bin Liu and HV Jagadish. 2009. A spreadsheet algebra for a direct data manipulation query interface. In *2009 IEEE 25th International Conference on Data Engineering*. IEEE, 417–428.
- [22] Douglas C Montgomery. 2017. *Design and analysis of experiments*. John Wiley & sons.
- [23] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.
- [24] Yogendra Patil. 2017. A Multi-interface VR Platform For Rehabilitation Research. *CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 154–159.
- [25] Yogendra Patil Paulo Lopez-Meyer, Stephen Tiffany and Edward Sazonov. 2013. Detection of cigarette smoke inhalations from respiratory signals using reduced feature set. *35th Annual International Conference of the IEEE EMBS*, 6031–6034.
- [26] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Equille, Maximilian Fabricius, and Srividya Subramanian. 2021. Balancing Familiarity and Curiosity in Data Exploration with Deep Reinforcement Learning. In *aiDM'21: Fourth Workshop in Exploiting AI Techniques for Data Management (ACM SIGMOD), Virtual Event, China, June 20-25, 2021*. 1–9.
- [27] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Equille, Maximilian Fabricius, and Srividya Subramanian. 2021. DORA The Explorer: Exploring Data with Interactive Deep Reinforcement Learning. In *The 30th ACM International Conference on Information and Knowledge Management, CIKM 2021, Online, November 1-5, 2021*. to appear.
- [28] Lilong Jiang Protiva Rahman and Arnab Nandi. 2020. Evaluating Interactive Data Systems - Survey and Case Studies. *The VLDB Journal*. 29 (2020), 119–146.
- [29] Daniel Vogel Quentin Roy, Futian Zhang. 2019. Automation Accuracy Is Good, but High Controllability May Be Better. *ACM Computer Human Interaction 2019 CHI* 2019, 1–8.
- [30] David A. Randell, Zhan Cui, and Anthony G. Cohn. 1992. A Spatial Logic Based on Regions and Connection. In *Proceedings of the Third International Conference on Principles of Knowledge Representation and Reasoning (KR'92)*. 165–176.
- [31] Arturo Realyvásquez-Vargas, Z Emigdio, Lilia-Cristina Morales, Jorge Luis García-Alcaraz, et al. 2020. Mental Workload Assessment and Its Effects on Middle and Senior Managers in Manufacturing Companies. In *Evaluating Mental Workload for Improved Workplace Performance*. IGI Global, 109–137.
- [32] Raphaëlle N Roy, Sylvie Charbonnier, Aurélie Campagne, and Stéphane Bonnet. 2016. Efficient mental workload estimation using task-independent EEG features. *Journal of neural engineering* 13, 2 (2016), 026019.
- [33] Yogendra Patil Stephen Tiffany, Edward Sazonov. 2014. Understanding smoking behavior using wearable sensors: Relative importance of various sensor modalities. *36th Annual International Conference of the IEEE EMBS*, 6899–6902.
- [34] Stefanie A. Wind and Yogendra J. Patil. 2018. Exploring Incomplete Rating Designs With Mokken Scale Analysis. *Educational and Psychological Measurement* 78, 2 (2018), 319–342. <https://doi.org/10.1177/0013164416675393> arXiv:<https://doi.org/10.1177/0013164416675393>