A Heuristic Approach for Dual Expert/End-User Evaluation of Guidance in Visual Analytics

Davide Ceneda (iD), Christopher Collins (iD), Mennatallah El-Assady (iD),
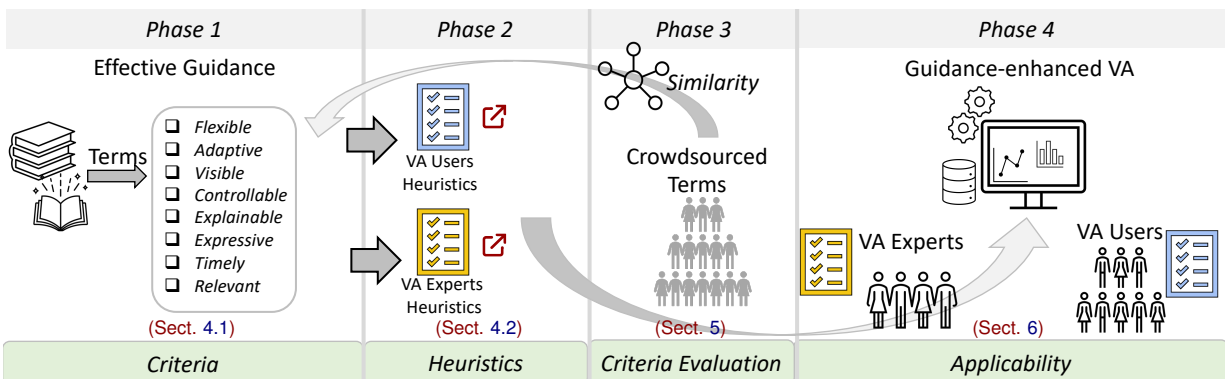Silvia Miksch (iD), Christian Tominski (iD) and Alessio Arleo (iD)

Figure 1: The methodological approach that guided the design of our evaluation methodology for guidance approaches.

Abstract—Guidance can support users during the exploration and analysis of complex data. Previous research focused on characterizing the theoretical aspects of guidance in visual analytics and implementing guidance in different scenarios. However, the evaluation of guidance-enhanced visual analytics solutions remains an open research question. We tackle this question by introducing and validating a practical evaluation methodology for guidance in visual analytics. We identify eight quality criteria to be fulfilled and collect expert feedback on their validity. To facilitate actual evaluation studies, we derive two sets of heuristics. The first set targets heuristic evaluations conducted by expert evaluators. The second set facilitates end-user studies where participants actually use a guidance-enhanced system. By following such a dual approach, the different quality criteria of guidance can be examined from two different perspectives, enhancing the overall value of evaluation studies. To test the practical utility of our methodology, we employ it in two studies to gain insight into the quality of two guidance-enhanced visual analytics solutions, one being a work-in-progress research prototype, and the other being a publicly available visualization recommender system. Based on these two evaluations, we derive good practices for conducting evaluations of guidance in visual analytics and identify pitfalls to be avoided during such studies.

Index Terms—Guidance, heuristics, evaluation, visual analytics.

◆

1 INTRODUCTION

Visual Analytics (VA) combines human and machine effort to generate insights [15, 29]. The problem, though, is that human-computer collaboration is typically imperfect due to noisy and incomplete communication on goals and tasks between user and system, which translates into time-consuming trial-and-error attempts at data analysis. For this reason, guidance approaches are studied to ease the user's analysis and close potential knowledge gaps that might hinder the analysis [7].

The literature on guidance approaches is rich and spans multiple fields [9]. Typical examples of guidance are recommender systems [21, 32, 35], which aim to suggest to the user how to analyze the data (e.g., the next action to take), or how to set up the visual environment to facilitate data exploration and completing analysis tasks. In general, guidance approaches can provide a variety of support to resolve knowledge gaps of different types (e.g., for supporting data manipulation [28], or to support the classification and modeling of data [13]) in different situations [19, 24, 27].

- *Davide Ceneda, Alessio Arleo and Silvia Miksch are with TU Wien.*
- *Christopher Collins is with Ontario Tech University.*
- *Mennatallah El-Assady is with ETH Zurich, AI Center.*
- *Christian Tominski is with the VAC Institute, University of Rostock.*

Besides practical implementations, guidance has also been studied extensively from a theoretical point of view [8, 14]. Recently, researchers made further steps to bridge this theory to practice, resulting in frameworks and guidelines for designing and implementing effective guidance [22, 45]. One of the open issues with guidance is that while existing frameworks and approaches claim that the envisioned guidance should be effective for the user (e.g., [5]), it is still not very clear how to evaluate in practical terms the extent to which the guidance succeeded in helping the user. While the literature concerning the evaluation of visual interfaces is abundant, the same cannot be said for the evaluation of guidance approaches. In other words, evaluating the effectiveness of guidance is an open challenge, with most existing works resorting to (sometimes inappropriate) methodologies initially created for evaluating visualizations, not guidance.

But how can guidance be *effective*? The term is rather vague when referring to guidance, and hence is open to many interpretations. Can we consider it effective when it improves the user's performance? Or when it is provided in a timely manner? As it is when evaluating visualizations, there are many facets to guidance effectiveness, too. If we aim to design and implement well-thought guidance approaches, we need to be able to specify with sufficient precision which qualities and characteristics a guidance approach should possess to work properly for the user. Conversely, this could also help practitioners understand why certain approaches fall short of providing effective support, which could lead to the design of better guidance in a broader sense.

To address these challenges, we propose a heuristic approach to eval-

uating guidance in VA. Building on a recent attempt at characterizing effective guidance design [5], we conducted extensive literature research and interviewed visualization experts to identify what characteristics make guidance effective, ultimately resulting in eight quality criteria. These were then instantiated into two sets of heuristics and questions that can be answered by *VA experts* (■ VAEs), such as designers, to assess the quality of the design, and by *VA users* (■ VAUs), to assess if the guidance was effective for them during the analysis when solving tasks – hence the "dual" nature of our methodology.

By evaluating guidance from the perspective of VAUs and VAEs, we aim to provide a comprehensive quality assessment of the guidance. Summarizing, the paper's contributions are as follows:

- A characterization of the *effectiveness* of guidance through the identification of a set of eight quality criteria (see Tab. 1, left column) that contribute to the value and to the effectiveness of guidance.

- Two sets of heuristics (see Tab. 1, center and right column) to evaluate and quantify the effectiveness of guidance, either from the perspective of VA experts (e.g., designers) or VA users.

- An in-depth evaluation of our methodology using two guidance-enhanced VA systems, and a qualitative analysis of the results.

- A detailed protocol and templates (download them here ⎋) to replicate the study and apply it to other guidance approaches.

2 RELATED WORK

In this section, we review the literature on the evaluation of visualizations and guidance.

2.1 Evaluating the Effectiveness of Visualizations

Studying the factors that contribute to the effectiveness of visualizations is a challenge tied to the development of evaluation methodologies. By evaluating them systematically, we can judge visual interfaces and get a glimpse of their potential and limitations. In other words, evaluation methodologies aim to assess the effectiveness of a given information display in supporting specific perceptual tasks [33].

The effectiveness of visual displays is the result of many different aspects concerning how the information is shown to the user. Miksch and Aigner argue that the *"effectiveness primarily considers not only the degree to which a visualization addresses the cognitive capabilities of the human visual system, but also the task at hand, the application background, and other context-related information, to obtain intuitively recognizable and interpretable visual representations"* [36, p.287]. These statements highlight why evaluating visualizations is a challenging task, and that accounts for the variety of domains and scenarios in which information visualization tools are used [44].

Among the many strategies to evaluate visualization, one of the most common is to measure the performance of multiple users using one or more tools in controlled experiments [41]. With this methodology, evaluators observe and collect metrics on how users perform data analyses and use a given tool. Typically, a baseline is used as a benchmark against which all the results are compared. Multiple different metrics can be utilized to assess the usability and performance of visual tools [11, 12]. For instance, execution time or error rates can be utilized to assess a visualization's effectiveness, i.e., the accuracy fostered by the visualization. Beyond accuracy, assessing effectiveness entails measuring the memorability of a visual display, i.e., recalling data cases or elements of the interface, as well as collecting qualitative and subjective aspects of visualization tools such as user's enjoyment, satisfaction, fun [42], and aesthetic pleasure [23].

An additional way for evaluating visualizations is using heuristics to assess usability and identify subjective and qualitative problems that might hinder the completion of the analysis [55]. The use of heuristics is common in human-computer interaction to test whether tools comply with well-known design principles and rules of thumb [17]. During the years, many heuristics have been identified to evaluate visualizations in specific contexts and domains [2, 18, 38, 43, 51, 54]. Molich and Nielsen identified nine principles and used them to evaluate visual interfaces, organizing usability issues according to their severity [37]. Heer et al. [25] identified a set of heuristics to evaluate the effectiveness

of animated transitions between well-known chart types. Mankoff et al. describe a set of heuristics for the evaluation of ambient displays, which are aesthetic visualizations portraying non-critical information on the periphery of a user's attention [34]. Wall et al. describe a set of heuristics to evaluate and quantify the value of visualizations, assessing their ability to communicate the *essence* of the data, the quality of the *insights*, the *confidence* of the user, and the performance and *time* of the analysis (i.e., the so-called ICE-T methodology) [50]. An interesting aspect of this methodology is that it provides a solid way to compare the qualities of multiple (even different) visualization approaches.

In summary, the effectiveness of visualization tools is the result of a variety of factors that mostly depend on how the visualization approach was designed and also on how people use and experience the tool. Taking inspiration from evaluation methodologies for visualizations, we aimed to identify what factors contribute to the effectiveness of guidance. In this aspect, our approach is similar to the ICE-T approach by Wall et al. [50] in that we aim to provide a clear numerical value to represent the effectiveness of guidance approaches and enable the comparison of their characteristics. Moreover, similar to the work by Stasko [48], we aim to characterize but also quantify the value of guidance in VA and the multiple facets of guidance effectiveness.

2.2 Evaluating the Effectiveness of Guidance

In parallel with researching visualizations, scientists also looked for approaches to support and enhance their use. In this regard, it makes sense to differentiate between *onboarding* approaches [49] – which are mostly utilized before the analysis (e.g., videos and tutorials) to facilitate their use and *guidance* approaches [7] – which are typically used during the analysis to enhance and facilitate the completion of tasks. In this paper, we focus on the evaluation of guidance techniques.

A guidance approach can be characterized by the knowledge gap(s) it aims to address, by the input used and output produced, and by the degree, which represents the amount of support given to the user [7]. Collins et al. expand the discussion of guidance and examine additional types of knowledge gaps [14]. Ceneda et al. review the literature concerning guidance approaches categorizing papers according to the guidance provided by the system to the user and according to the guidance the system receives from the user [9]. Sperrle et al. expand this categorization, focusing on learning and teaching processes [46]. Han and Schulz investigate a methodology for implementing guidance based on decision support theory [22].

When analyzing the literature on guidance approaches, we can see no single methodology for their evaluation. Even more, the evaluation of guidance and visualization tools are typically carried out together. Hence, the effects and the contribution of guidance to the overall analysis process are not immediately clear or recognizable. Heer et al. describe an approach to enhance the exploration of communities in large networks [24]. To evaluate the approach, the authors focus on its usability [38], following a standard methodology to evaluate visualizations. When evaluating guidance, most approaches focus on assessing whether the introduction of guidance leads to performance improvements. For instance, Bouali et al. present a tool that can recommend appropriate visual encodings to the user; the approach is evaluated in a controlled experiment comparing the performance of users receiving guidance with those applying a completely manual workflow [4]. Gotz et al. also consider performance improvements (i.e., (task) completion time and error rate) to evaluate their guidance-enhanced approach to suggest appropriate visualizations based on user interaction trails [20]. May et al. describe an approach to guide the exploration of large graphs using signposts. To evaluate it, the authors utilize completion time and two click-based metrics, showing how guidance reduces them significantly. Wongsuphasawat et al. focus on the effects of guidance and report how the use of a recommender system encourages the use of visualizations of interest [53].

Beyond raw performance, Ceneda et al. explore the effects of guidance on users, describing how, in particular, the guidance provided changed their exploration strategies, i.e., the way users analyzed the data and how, conversely, this led them to a deeper understanding of the data and increased confidence towards the results obtained [10].

Krause et al. assess how their guidance-enhanced approach can foster deeper insights into predictive models and improve model interpretability [30]. An interesting way to evaluate guidance is presented by Krishnamoorthy and Brusilovsky, who presented a tool to guide students toward appropriate examples of a pre-selected topic [31]. In particular, the authors focus not only on the evaluation of the interface design, but also on measuring the amount of help provided to the users, their ability to perceive progress, the navigability of the tool, and on measuring whether a certain goal was achieved thanks to the guidance. O'Donovan et al. propose a qualitative evaluation of their guidance-enhanced approach to improve visualization design with layout suggestions and highlight how the designs recommended by the guidance system were favorably evaluated by experts [39].

We emphasize that similar to what happens for visualizations, guidance approaches are applied in a variety of scenarios providing diverse types of support. Consequently, no single way to evaluate guidance is known in the literature. Some approaches use mere performance as an indicator of the effectiveness of guidance. Others check whether an approach complies with suitable design practices, such as usability, recall, and interpretability of the provided support, assessing guidance from a designer's perspective. Finally, others are instead concerned with evaluating the qualitative aspects of guidance and its influence on the analysis, evaluating the provided guidance from a user perspective.

Drawing from the literature, our goal is to derive an evaluation methodology that could shed light on the effectiveness of guidance, taking into account these multiple perspectives and providing a comprehensive assessment and a common ground for the comparison of multiple types of guidance. For this reason, we go a step further than existing approaches, describing two sets of heuristics to evaluate guidance from the perspective of VAEs and VAUs. The first step in the process to derive our methodology was to identify the factors (i.e., a set of quality criteria) that contribute to making guidance effective. In the following, we describe this process and its outcome.

3 METHODOLOGY

We followed an iterative process (see Fig. 1), inspired by He et al. [23], to derive our heuristic approach in four phases: (i) Identification of quality **Criteria** (see Sect. 4.1). (ii) Instantiation of the **Heuristics** (see Sect. 4.2). (iii) **Criteria Evaluation** (see Sect. 5). (iv) **Applicability** of the heuristics (see Sect. 6).

We began our research by conducting a literature review to identify what criteria were used in the literature to evaluate visualizations and guidance approaches. Our goal was to pinpoint the terms scientists associated most often with the effectiveness of an approach. In this initial phase, we collected a comprehensive set of terms, including terms like trustworthy, efficient, appropriate, and accessible. Afterward, we organized the terms in a similarity graph, where terms were represented as nodes and their similarity was shown by edges, to identify overlaps among similar terms relating to potential effectiveness criteria and evaluation methodologies. We used the graph to filter down the initial set of quality criteria and organized them into categories, structuring them around the requirements for guidance design by Ceneda et al. [5]. In the following phase, we reviewed the criteria and refined the structure of items and terms in a collaborative manner. This process led to the identification of eight quality criteria that contribute to the effectiveness of guidance. We instantiated these criteria in two sets of heuristics for use by VAUs and VAEs to evaluate guidance.

We conducted three evaluations of our heuristic approach. The first evaluation was focused on evaluating the criteria. We sent out a questionnaire to more than 80 VA scholars with experience in working with guidance and designing visualizations. We received answers from 25 of them, whose feedback helped us refine the criteria and their characterization and provided us with evidence of the soundness of our methodology. The focus of the second and third evaluations was on the applicability of our approach in practical scenarios. We first asked a group of VAEs to apply the heuristics to evaluate existing guidance approaches. Second, a group of VAUs worked with a guidance-enhanced tool and subsequently filled out our heuristics-based questionnaire to evaluate the effectiveness of the guidance they received.

4 THE EFFECTIVENESS OF GUIDANCE

The first step in building an evaluation approach for guidance was identifying the factors and the criteria contributing to guidance effectiveness on which to ground the methodology. A preliminary set of criteria was gathered by reviewing the literature.

Literature review   We searched the proceedings of major conferences as *Conference on Human Factors in Computing Systems (CHI)*, *IEEE VIS* and *EuroVis*, and journals like *TVCG*, *CGF*, using terms like "guidance," "visualization," "evaluation," as well as synonyms and related words. We identified more than a hundred papers dealing with visualizations, evaluation methodologies, and heuristics and a slightly smaller number of papers dealing with guidance. In particular, we considered literature reports and surveys (see Sect. 2) from which we extracted terms and criteria and categorized papers according to whether they discussed "metrics" or "heuristics". For guidance-enhanced approaches, we started our research from the survey by Ceneda et al. [9], which we expanded to consider more recent works. Since we could not ground our research on existing evaluation methodologies for guidance, we instead extracted terms and derived qualitative criteria directly from the text, also analyzing the reported user feedback. Typically, guidance approaches are evaluated using qualitative measures. Studies often include direct feedback and citations of the participants (e.g., "*thanks to the guidance, the participant reported an increased trust in the results obtained*" [10, p.7]). In these cases, we synthesized participants' feedback, organizing it in tokens (e.g., "increased trust") representing different facets of guidance effectiveness. From this analysis, we collected more than 180 terms comprising heuristics, metrics, and quality criteria, of which around 40 were specifically used for evaluating guidance-enhanced approaches. We did not immediately discard qualities used solely to characterize the effectiveness of visualization approaches. Instead, we kept them to understand whether they could also be applied to guidance, given the tight relationship between the two.

Figure 2: The similarity graph of terms, used to derive a minimal set of quality criteria. Nodes represent qualities and edges represent the similarity between concepts. The full graph is provided as supplementary material.

Organizing and structuring   Next, we worked to give structure to the terms collected and organized them in a graph. This graph revealed similarities that allowed us to reduce the number of terms to a manageable level. In 2020, Ceneda et al. [5] presented a framework to support the design of guidance approaches comprising steps, questions, and threats that designers must address. Their framework is supported by a list of requirements representing ideal qualities that a guidance approach should have in order to effectively address the users' knowledge gaps. According to this previous work, effective guidance derives from the following qualities: *Adaptive*, *Available*, *Trustworthy*, *Controllable*, and *Non-disruptive* [5]. We used these design requirements as initial aggregation nodes for the terms we collected reviewing the literature. We used the online visual collaboration platform Miro to create a similarity graph (with 70+ nodes and 90+ edges). Fig. 2 shows a portion of the graph (the complete similarity graph is available as supplementary material). Nodes (in blue) represent the terms and the qualities we collected, which we organized around the requirements described by Ceneda et al. (in orange). Edges between nodes represent possible analogies between concepts. We used this graph to refine and filter the list of terms, equally in a top-down and bottom-up manner, but

also to capture the different nuances of guidance effectiveness. For example, thanks to this structuring, we realized that the timeliness of the guidance and the expressiveness of its encoding contribute to its non-disruptiveness. Similarly, we found that the trustworthiness of guidance depends on its explainability and, therefore, merged the two qualities. The result of this phase was a set of 15 terms and qualities, which we further examined in the following phase.

Refining the criteria   While the previous phase was run mainly by two authors, the whole team collaboratively reviewed and examined further the resulting set of terms and qualities. In particular, additional similarities among terms were identified and some of them (e.g., non-disruptiveness) were deemed misleading, and thus, deleted. Hence, the terms were additionally aggregated, and the different nuances related to each quality were captured and formalized in a set of eight quality criteria capturing the effectiveness of guidance. Each criterion was, in the end, represented (or summarized) by a single term. To avoid confusion and potential ambiguities, we also provide a definition of each term.

4.1   Quality Criteria and Definitions

Here, we summarize our definitions and characterize the qualities that best describe the effectiveness of guidance. Note that what follows already includes the changes derived from the expert feedback gathered during the evaluation of our methodology (see Sect. 5).

After listing their definitions, we characterize each quality in more detail and refer to relevant literature in separate paragraphs.

- **Flexible** – The guidance is flexible when the degree of support adequately adapts to the analysis situation(s).

- **Adaptive** – The guidance is adaptive when it considers user preferences, habits, and current task requirements.

- **Visible** – The guidance is visible when its recommendations and status can be clearly distinguished in the visual environment.

- **Controllable** – The guidance is controllable when the user can switch the guidance on/off, request alternative recommendations, or revert previously followed recommendations.

- **Explainable** – The guidance is explainable when its recommendations are easily understood and the way they were generated is made transparent to the user (also on demand).

- **Expressive** – The guidance is expressive when its encoding is appropriate and users can extract the information needed to make analytic progress.

- **Timely** – The guidance is timely when it is provided on time and only when needed.

- **Relevant** – The guidance is relevant when it guides users toward their analytical goals and supports the completion of tasks.

Flexible   The extent and type of support the user receives during the analysis (i.e., the *guidance degree* [7]) are dynamically adjusted. This criterion explicitly addresses the need to identify a suitable guidance degree and the extent to which the system can adjust the amount and the type of support during the analysis (i.e., how tailored and stringent the guidance within a user task). Typically, the amount of guidance the user needs varies during the analysis, so the system must react to it dynamically to support the user effectively.

Adaptive   The content of the guidance (i.e., the information provided to the users to guide them) is adapted according to the analysis context, for example, when the user task or preferences change [45]. While the previously-described flexibility criterion addresses the need of identifying the appropriate *degree* of guidance, this one addresses the need to adapt the *content* of the suggestions given to the user. Typically, adaptiveness can be achieved by considering analytic provenance information, user preferences and user experiences, habits, and interactions with previous suggestions (e.g., rejected guidance) [43], which can all be used to generate user-tailored content.

Visible   This criterion ensures that the status of the guidance (e.g., if it is active or not, and its parameters) is visible to the user using appropriate visual feedback and within reasonable time [43]. That is, the guidance is effective if the user can perceive it. Similar criteria are typically used also for visualizations (e.g., [38]). In this, we extend their application to guidance-enhanced approaches as well.

Controllable   It must be possible for the user to fine-tune and control the parameters that influence the way the guidance is produced [5]. This criterion aims to evaluate whether a guidance approach allows the user to control, for instance, the level of details and the amount of guidance provided (see *flexibile*), and switch it off occasionally. Drawing from well-known heuristic approaches, an effective guidance system should show marked exits, support undo/redo of actions [37], allow easy access to parameters and settings [38] and, more in general, to orientation features, for improving usability and hence, guidance effectiveness.

Expressive   This criterion aims to assess whether guidance suggestions can be provided and communicated to the user using an appropriate language (i.e., using the user's language [37]). The chosen encoding should be used consistently throughout the system so to avoid confusion and ambiguities and to make the guidance recognizable in the visual interface (compare, for instance, the *legibility* heuristic [43]). The expressiveness of guidance also concurs with its explainability.

Explainable   This criterion aims to assess the way the guidance is communicated and understood by the user. Explainable suggestions concur to increased acceptance and trust (see guidance trustworthiness [5]) towards the received guidance. Furthermore, to uphold effectiveness, the explainability should not be limited to the suggestion itself but also clearly expose the process that led the system to it (which is typical in ML applications [47]). The user should be entitled to ask for more explanations and details, if necessary (see the *controllable* criterion). Generally, the guidance should adhere to the principle of least astonishment [3], i.e., the system should behave in a way that most users will expect. When this does not happen (e.g., when the user has to be addressed toward a new analysis path), each action and suggestion coming from the system should be adequately explained to avoid confusion.

Timely   The timeliness criterion assures that the guidance is provided at the right time, without delays, so as not to disturb or interrupt the user, or interfere with other tasks (e.g., when the task changes and the guidance is late and, hence, not relevant anymore). Together with the use of appropriate explanations, the timeliness of guidance concurs to make it accepted and trusted by the user [6].

Relevant   The last criterion deals with assessing whether the guidance can guide the user toward relevant analysis results and, at the same time, avoid that the user makes errors in the process. The guidance should have an aim and be able to perform basic tasks to support the user. The relevance of guidance is related to its ability to achieve these aims and perform these tasks. Typically, the system should be able either *to guide* users toward the analytical goal, to *correct* and readdress them along other analytical paths if problems are detected, and *to ask* for user input if needed (see the *guide*, *correct*, and *ask* tasks of Pérez-Messina et al. [40]). The analysis process is typically complex, and due to this complexity, many mistakes can happen. Drawing from accuracy metrics, the system should help users avoid analysis errors, or at least help them recover and get back on track, exposing uncertainties and biases [2]. In this regard, the combination of error detection and prevention functionalities accompanied by appropriate explanations (see *explainable* and *expressive* criteria) is necessary to ensure the user does not return to the erroneous analysis path. The guidance should also lead users towards their analysis goals, help them discover the unexpected when possible, or even generate hypotheses about the data in the first place. If a goal cannot be immediately identified (e.g., at the beginning of the analysis, goals could be undefined), the system can poke the user for details about the scope and goals (see the "ask" tasks [40, 45]). All these results could be achieved in multiple ways. Hence, the system should support the exploration of alternatives when several actions are possible depending on the analysis context. The ability of the system to achieve these guidance tasks contributes to its relevance.

Effectiveness Criteria	Heuristics for Visual Analytics Users (VAUs)	Heuristics for Visual Analytics Expert (VAEs)
Flexible — the guidance degree is adequate and adapts to the situation	UH1 The guidance I received matched my need for support to resolve the problems I encountered	EH1 Make the guidance degree dynamic
Adaptive — the guidance system considers user's preferences, habits, and the current task requirements for generating suggestions	UH2 When my task and goals changed, so did the guidance I received; UH3 I felt the system was considering my input (e.g., previous selected suggestions/interactions) and/or personal preferences when producing the guidance	EH2 Use different sources of (user) input, e.g., selections, and past interactions to fine-tune the guidance
Visible — the guidance and its status are easy to identify and well integrated with the rest of the visual environment	UH4 I could easily identify the guidance and distinguish it from the rest of the visual analytics environment	EH3 Integrate the guidance in the interactive visual environment
	UH5 The status and parameters of the guidance (e.g., if it is switched on, or if it is working, its parameters) were always visible on the screen	EH4 Employ a meaningful visual schema to represent the status of the guidance and its parameters (show internal status)
Controllable — the guidance can be steered, allow the user to evaluate alternatives, switch on/off, revert the guidance	UH6 I could explicitly control the guidance by requesting or disabling it at any time	EH5 Allow requesting guidance on demand
		EH6 Allow switching the guidance off
	UH7 The system allowed me to assess alternative suggestions and compare them, (e.g., see the results of alternative) suggestions without the need to apply them (foreshadowing)	EH7 Facilitate assessing alternative suggestions (minimal memory load, minimal actions)
	UH8 I could directly provide feedback on the guidance I received	EH8 Allow direct feedback to the guidance suggestions, (e.g., rating them) (immediate feedback)
	UH9 I could fine-tune the guidance settings/parameters (e.g., change amount of suggestions, formulate queries)	EH9 Facilitate the specification/modification of the guidance parameters (e.g., specify queries)
Explainable — the guidance and the way it was generated can be easily understood	UH10 I understood what results the system was guiding me to; UH11 The guidance suggestions were easy to understand	EH10 Display sufficient information to facilitate understanding of guidance (sufficient information)
	UH12 I understood the reason why the suggestions were provided to me (e.g., because of how I interacted with the tool in the past); UH13 I could prompt the system for (more) details regarding the guidance, (e.g., ask for explanations); UH14 I could trust the guidance	EH11 Provide details-on-demand for explanations (easy transition to more in-depth information)
Expressive — the visual encoding of guidance is appropriate	UH15 The guidance suggestions used a language I could understand	EH12 Use the user's language to provide guidance
		EH13 Use an intuitive and consistent visual encoding for guidance (consistency)
	UH16 The guidance encoding is unambiguous, (i.e., not open to more than one interpretation)	EH14 Avoid misleading representations for the guidance (avoid ambiguity)
Timely — the guidance is provided on time when needed	UH17 The guidance was provided at the right time when I needed it; UH18 The guidance did not disturb my workflow	EH15 Deliver guidance in real-time, without delays for the user
Relevant — the guidance helps prevent or recover from errors, guides towards a goal, helps discover the unexpected, provides multiple options and perspectives	UH19 The guidance helped me get out of a stalled analysis	EH16 The system prompts guidance to recover from error states (e.g., empty search results) and stalled analyses
	UH20 The guidance helped me solve my tasks in an accurate way, reducing the possibility of doing errors	EH17 The system provides guidance to warn of potential inconsistencies or errors in analyses
	UH21 The guidance saved me time; UH22 The guidance facilitated my reasoning; UH23 The guidance helped me answer questions about the data; UH24 The guidance was adequate for the task I was performing; UH25 Through guidance I made serendipitous discoveries; UH26 The guidance helped me generate new hypotheses about the data; UH27 I felt confident about the results obtained with guidance support	EH18 Help explore all alternatives when several actions are possible depending on the contexts

Table 1: The sets of heuristics to evaluate the effectiveness of guidance. The columns list the quality criteria associated with the effectiveness of guidance (left), the VAUs heuristics (center), and the VAEs heuristics (right). Rows specify the actual heuristics.

4.2 Heuristics for Dual Expert/End-User Evaluation

We derived the heuristics from our quality criteria such that the resulting heuristics covered all the nuances specified by the criterias' definitions. The heuristics are listed in the table in Tab. 1 in the ▨ central and ▨ right columns along with the quality criteria in the left column.

Evaluators To whom is our methodology addressed? During the initial phases of research, we realized that certain aspects of guidance could be better judged after having experienced the tool and received the guidance in the first place. For example, to check if the guidance was provided on time or helped in getting out of problematic situations, it makes sense to refer to VA users who actually received the guidance. Likewise, we realized that VA experts (e.g., designers) could better assess other criteria from the first stages of development without needing to actually experience the tool. This led us to derive and instantiate the criteria in not one but two sets of heuristics, each focused and tailored for a specific type of user: VAEs and VAUs.

In total, we describe a different number of heuristics, 27 to be assessed by VAUs and 18 for VAEs. We expect the expert evaluators to be knowledgeable of the guidance concept and of visualization design in general. We do not expect the user evaluators to be knowledgeable besides being able to use a VA system.

Formulation We formulated the heuristics keeping in mind their specific audience. For instance, this implied formulating the VAUs heuristics using the first-person singular and avoiding domain-specific terms and references to guidance terminology (e.g., references to the guidance degree or the knowledge gap were avoided), as we do not expect VAUs to possess any specific knowledge of those terms. Also, VAUs heuristics were formulated positively using single sentences and avoiding conjunctions, to ease their interpretation and their assessment in a survey. Conversely, heuristics meant for VAEs are shorter and typically formulated as rules of action.

4.2.1 Running the evaluation

Our heuristics are to be applied practically using different questionnaires for VAEs and VAUs. While VAUs need to have direct access to the studied system and its guidance to evaluate the effectiveness, VAEs do not necessarily need to use the tested tool (i.e., the system could still be in the design phase) but may rely instead on design descriptions (e.g., sketches, videos, mockups).

Given the different target audiences of the two heuristics sets, we expect them to be rated differently according to the following scheme.

VA experts VAEs can rate the system's adherence to the heuristics assessing the application or the violations of the different design principles (e.g., using a Likert scale from 1–clearly violated to 7–clearly applied, or N/A not applicable). The identified violations of design principles and heuristics can be reported. Subsequently, severity ratings [38] (e.g., from 0 – not a problem, to 4 – design catastrophe) could be used to assess their severity and allocate enough resources to possibly solve them to improve guidance effectiveness. In line with Molich and Nielsen, we expect VAEs heuristics to be rated at least by 3-5 evaluators to identify a wide range of design problems. Finally, given the way we created the heuristics, we assume that the evaluation of guidance can be run in parallel with the typical assessment of usability problems [37].

VA users On the other hand, as VAUs heuristics are formulated as statements, we expect VAUs to rate them using a 7-point Likert scale ranging from 1-strongly disagree to 7-strongly agree or N/A not applicable, to assess their agreement with the reported statement/heuristics. Subscores can be calculated for each criterion by summing the agreement (1-7), representing an estimation of the different qualities of the system under analysis. A final score can be obtained by summing the subscores to obtain an overall estimation of the effectiveness of guidance. Calculating and using scores and subscores enables the direct comparison of (different) guidance-enhanced approaches and their characteristics. Finally, VAUs heuristics should be evaluated with at least 15 users in order to achieve a reliable result.

Questionnaires The questionnaires for VAEs and VAUs, and the protocol to run the evaluations can be found in this Git repository ☐ [1].

5 CONSOLIDATING QUALITY CRITERIA

As part of our effort to assess the quality, clarity, and comprehensiveness of our set of effective guidance criteria, we ran an expert evaluation with a twofold purpose. First, we wanted to gather feedback about the selection of terms, on the lookout for what we might have missed during the literature review (see Sect. 3). Second, we were interested in the participants' opinions about the quality criteria and their definitions, looking for alternate formulations and/or additional quality criteria. Here, we describe the study design that led us to those results.

Procedure The evaluation was arranged as a two-part online survey. After completing the informed consent form, we described the scope of the study to the participants. In the first part of the study, we asked the participants to think of at least three adjectives or qualities they associated with their personal idea of effective guidance. We also asked, as an optional question, to provide and describe additional qualities, discuss examples, or further justify the given adjectives. In the second part, we presented visualization experts with nine quality criteria with their proposed definitions (different from the final list presented in Sect. 4.1 and available in the supplemental material). For each one of them, we first asked the participants their opinion about their relevance in the context of a VA system on a 1 (not important at all) to 5 (essential) scale, e.g., "*In the context of a VA system, in your opinion how relevant is it for the provided guidance to be adaptive?*". The participants were also asked to evaluate our definitions of the criteria, stating if they agreed with them or not. Participants were encouraged to provide their own definitions and additional comments about their assessment. Finally, participants could provide overall feedback on the set of criteria as a whole, evaluating, for instance, the importance of each criterion compared to the others in capturing the effectiveness of guidance.

Participants We administered an online questionnaire to scholars with proven research experience (i.e., a publication record) in visualization and guidance. We contacted them by email and gave them three weeks to complete the survey. Participants were not remunerated for their participation. Ultimately, 25 of them answered the questionnaire.

Outcome From the answers, we gathered around 100 terms, qualities, and adjectives that experts associated with effective guidance. We preprocessed the list to remove duplicates and merge similar terms. This step was facilitated by the examples and explanations provided by the participants. Overall, we obtained around 50 terms to merge with those collected from our initial literature review (see Sect. 3). For this purpose, we created a similarity graph in which the crowdsourced qualities (nodes) were associated (with edges) with previously gathered literature terms according to their similarity.

The similarity graph helped us refine the criteria and capture nuances we did not previously consider. For example, participants reported confusion when reading the definitions of *adaptive* and *flexible* guidance and suggested exchanging their names. Other fine-grained adjustments were made to the other criteria and their definitions. For example, based on the feedback received, we renamed a criterion to *relevant* (which previously was referred to as *purposeful*) and merged it with one concerning the *resiliency* of guidance. Finally, it is worth mentioning that the study participants mostly associated guidance effectiveness with its *explainability* (i.e., among the reported terms, many cited "clear," "transparent," "descriptive,") and its *relevance* (i.e., "successful," "insights," "task-oriented," "helpful," "meaningful"). This information can, for example, be helpful for weighting the severity ratings in actual evaluations of guidance systems.

All the feedback and insights we gathered from this first study went into the refinement of our quality criteria and respective definitions (see Sect. 4.1).

6 EVALUATING HEURISTICS IN PRACTICE

Having formalized criteria after our first expert evaluation, we aim to evaluate how our heuristics perform in practice in assessing the guidance quality of existing VA systems or in providing insights that can contribute to their design. For this purpose, we designed and ran two more studies, one involving 5 VAEs and another with 39 VAUs, described in the following.
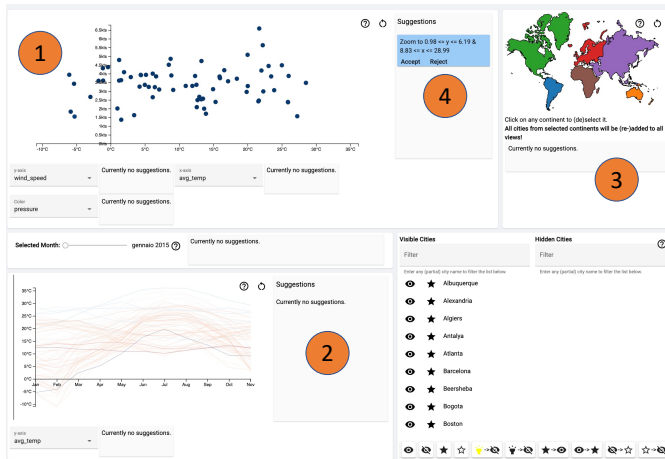
Figure 3: The guidance-enhanced system we used for the expert evaluation. It features different views, including a scatterplot (1), a line chart (2), and a map (3), as well as the possibility to provide guidance to the user. The guidance appears next to each chart as a blue textual box (see for example (4) close to the scatterplot), suggesting to the user what parameters and what data to select in the given visualization.

6.1 🟨 VAEs Evaluation

The goal of this expert evaluation was to determine whether VAEs could apply our heuristics to assess the guidance quality in the design stage.

6.1.1 Study design

According to Nielsen [38], it is often the case that evaluators do not have access to the actual system when evaluating the ongoing design, especially if the evaluation takes place early in the development cycle. Hence, to recreate the conditions of a realistic evaluation of a guidance-enhanced VA prototype, we assembled a video of a system under development, which we then showed to the study participants.

The video showed six scenarios of how later users are supposed to use and benefit from the guidance features included in the system. The system in question is currently in development as part of another research project involving some co-authors of this paper. It can provide different types of guidance at different stages of the data exploration process, and is portrayed in Fig. 3.

The system employs multiple coordinated views to support various analytical tasks. In our case, it is used to analyze a dataset comprising weather data collected from different cities worldwide. The task of the user is to identify a city with a specific climate given a set of constraints. The data are shown in a scatterplot, a line chart, and a world map, respectively labeled as (1-3) in Fig. 3. Using these views, the user can choose to visualize, filter, and select different data dimensions and analyze their evolution over time. Guidance is provided as blue textual boxes that appear next to the target chart (e.g., see (4) in Fig. 3) suggesting the user a specific action (e.g., change of axis, value filtering, etc.) on that chart. By hovering over the suggestion, a preview of its effect is shown to the user.

When the video for the experiment was recorded, guidance could be given directly by a hidden "wizard" user, which allowed us to stage different guidance quality criteria from our list using the system features and record it from the user's perspective. The rationale is to make sure that VAEs received guidance with specific characteristics so that we could check if and to what extent they could identify them.

All scenarios show the user interacting with the system while solving a task. For the purpose of our study, the guidance was fine-tuned as follows:

F1. In all scenarios, guidance is not explained, i.e., no reason is given as to why a suggestion popped up (*un*-explainable guidance).

F2. In one scenario, some suggestions were provided with some consistent delay to the user on purpose (*un*-timely guidance).

F3. In a different scenario, the system provided irrelevant guidance,

i.e., suggestions that were not aligned with the user task (*irrelevant* guidance).

F4. In another, the system provided guidance to correct and refine the user selection (*adaptive* guidance).

F5. In general, the system could provide, for each view, different degrees of guidance, such as *directing guidance* and *orienting guidance*, as well as no guidance (see the guidance degrees described in [7]) and the guidance type varied during the analysis (*flexible* guidance).

Participants We contacted 9 VAEs by email for the expert evaluation. They are all Ph.D. students or already hold a Ph.D. degree in visualization, and hence can be considered experts. Eventually, 5 VAEs participated in our study.

Procedure The participants received a link to an online survey by email. After signing a consent form, the participants received a first introduction to the analysis tool. They were given a textual description and some static figures of the system and of the different views. Furthermore, we asked them to watch a one-minute video showcasing the main features of the system, how to use them, and the various forms of supported guidance. After asking if there were questions about the system, we show the video with the six scenarios to the participants in the next phase of the evaluation. The participants could freely watch the video and the different scenarios as often as they wanted. Participants could also go back to the video when answering the questionnaire if they missed some detail. The video lasted 5 minutes; it had no audio but displayed a few captions describing the user's actions or the start of a new scenario. At the end of the video, we asked the participants to fill out a questionnaire containing the expert heuristics (VAEs) listed in Tab. 1. The participants rated the heuristics following the protocol described in Sect. 4.2.1.

	Adaptive	Visible	Timely	Expressive	Flexible	Relevant	Controllable	Explainable
E1	7	7	7	6	7	6	5,4	5,5
E2	7	6	7	5,67	6	5,67	4,4	4
E3	6	6	6	5	5	4	3,8	3,5
E4	6	5,5	6	5	5	3,67	3,2	2,5
E5	5	5,5	3	4,33	2	3,67	2,2	2,5
AVG	6,2	6	5,8	5,2	5	4,6	3,8	3,6

Figure 4: Results of the experts (E1-E5) evaluation. In the table, the ratings are grouped according to the different effectiveness criteria and ordered based on the ratings. Reddish cells represent ratings close to 1 (heuristics violated), while greenish cells represent ratings close to 7 (heuristics applied). Columns are ordered based on the average rating, from higher to lower ratings.

6.1.2 Outcome of the expert evaluation

All five experts completed the survey and the results were used for our evaluation. The results are summarized in Fig. 4. For convenience, we grouped the answers according to the quality criteria and calculated an average of the Likert ratings. The system obtained a total average score of 5.1 (of possible 7). In the context of our evaluation methodology, an average score equal to or above 5 means that the system provides *sufficient* guidance [50].

In general, the expert participants were able to apply our heuristics to identify and rate all the characteristics of the guidance, in line with the features (F1–F5) described earlier. The participants identified the guidance degrees (F5) provided by the system and assessed its *Flexibility*, but also sometimes highlighted the lack of dynamism in adjusting the guidance degree to the user needs. The *Adaptability* of the guidance was positively assessed: The participants recognized that the content of the suggestions was adapted according to the user needs and that the

system possibly considered analytical provenance information to generate content. The guidance also received favorable ratings for its *Visibility*, as it was always easy to spot newly generated suggestions. At the same time, VAEs found that the system did not provide access (i.e., Visibility) to the guidance parameters and settings. The *Controllability* of the guidance received bad ratings overall. The participants complained that the user was not allowed to ask for guidance, modify its parameters, or turn it off. Similarly, VAEs recognized that the guidance did not have explanations (F1), nor could it be requested by the user, so its rationale or motivation could only be guessed. The *Expressiveness* received an average rating of 5.2, which is greater than five and a weakly satisfactory score. However, the participants highlighted how the encoding of the guidance, which sometimes employed screen coordinates in pixels for the suggestions, was not particularly user-friendly and inexpressive and suggested changing it. The *Timeliness* of the suggestions received good ratings, with one expert recognizing that, in one scenario, the guidance was delayed (F2). Finally, the *Relevance* of the guidance received mixed ratings. Some participants recognized that the guidance was typically on-spot (F4). However, they also recognized the presence of misguidance (F3) and that the missing explanations exacerbated this situation, which made them decrease the final rating.

6.2 ▪ VAUs Evaluation

We set up a second study to assess the applicability of our heuristics for VAUs. While the goal of the expert evaluation was to assess if a small number of experts could identify a significant number of design issues, our goal for the user evaluation was to assess the coherence of the provided answers and if they could be used to shed light on the effectiveness of a guidance-enhanced system.

6.2.1 Study design

As specified in Sect. 4.2.1, we assume users are exposed to our heuristics only after experiencing the guidance to judge its effectiveness. For the study, we employed the open-source guidance system "Voyager" (see Fig. 5), a mixed-initiative system that enables users to perform a guided exploration of a dataset [52]. In terms of guidance, the system offers different kinds of recommendations to support the selection of appropriate visualizations to perform the analysis. Upon the user selecting the data dimensions to analyze, the system generates a list of appropriate visual encodings based on their perceptual properties and statistical measures of the dataset. Moreover, when the user selects a specific visualization, the system can recommend additional data dimensions to be explored as well as alternative encodings of the same data.
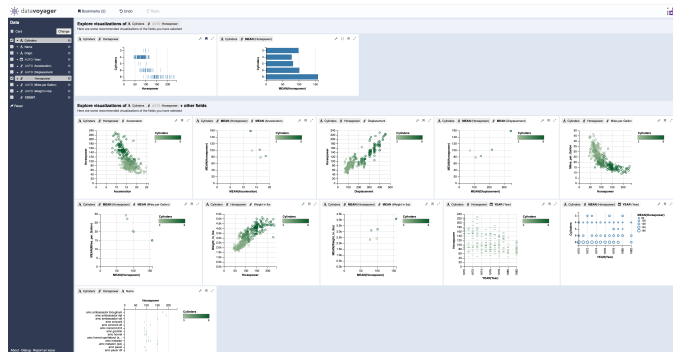
Figure 5: Voyager [52], the guidance-enhanced system for our user evaluation.

The Voyager system can be categorized as providing *orienting* and *directing* guidance [7]. Orientation is upheld by presenting to the user a broad set of charts to choose from to start the exploration as soon as the data is selected. Directions are given when the user is looking for alternative encodings, to which the system responds by providing a ranked list of different possibilities. Moreover, the system features (to a certain extent) *adaptiveness*, in that the process of generating suggestions considers the user input. Voyager guidance has no *explainability*

in its design and implementation, and the timeliness of the recommendations as well as means to control the guidance parameters were not considered during the system design. However, the authors of Voyager found out that the guidance promoted increased coverage of the data variables under analysis (thus, it was deemed *relevant*).

Participants We asked 50 students to take part in this evaluation. They were part of a bachelor-level course on information design and visualization and have some preliminary knowledge of visualization concepts, but cannot be considered VAEs. Students could freely decide to participate either in our study or take part in an alternative assignment.

Procedure After signing a consent form, the participants were guided through a practical tutorial to make them familiar with the system features and the guidance, using the embedded *car* dataset. Afterward, they were asked to switch to the embedded *FAA Wildlife Strike* dataset – which includes statistics of animal striking aircraft – to solve an analysis task. The task comprised getting familiar with the dataset and "getting a comprehensive sense of the dataset's content, collecting interesting patterns, trends, or other insights worth sharing with colleagues." We did not impose a strict time to complete the task, but we suggested using the tool for 5 minutes at least. To motivate them to interact with the system, we asked them to write a short report (a few lines) of their findings at the end of the study. Finally, after the task was completed, we administered a questionnaire containing our 27 VAUs heuristics.
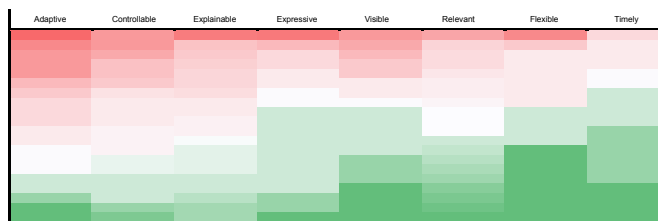

Figure 6: Results of the user evaluation of Voyager [52]. For space reasons, we do not report the ratings but simply color-coded the table according to the ratings. Reddish cells represent ratings close to 1 (disagree with heuristics), while greenish cells represent ratings close to 7 (agree with heuristics). Columns are ordered based on the average rating, from lower to higher.

6.2.2 Outcome of the user evaluation

In total, 39 students answered to our questionnaire. All the received answers were deemed valid (i.e., no incomplete questionnaire was submitted) and valuable for our evaluation. The results are summarized in Fig. 6 (the full graph is provided as supplementary material), in which results are grouped according to the quality criteria. To unfold the evaluation outcome, we provide a high-level assessment of these results. Moreover, we conduct a comparative analysis by running the same protocol but with VAEs participants, albeit using expert heuristics.

Qualitative analysis The results collected provide an assessment of the Voyager system from a guidance perspective. Participants assigned low ratings to the *Adaptability* of the guidance. While the system can provide guidance based on the currently selected data, the participants assessed that the guidance was not adapting to changing needs (UH2, central column, Tab.1) nor considered past interactions and analytic provenance to update suggestions (UH3). Also, the *Controllability* criterion received mixed evaluations. The participants highlighted that the system does not allow the user to provide feedback to the guidance (UH8) or fine-tune it (UH9). The study participants positively assessed its *Explainability*, as they could easily understand why particular suggestions were provided (UH12). For this reason, they also trusted the guidance (UH13) but highlighted how additional explanations could not be requested on demand (UH14). Finally, analyzing the *Relevance* of the guidance, the participants reported that they were not supported in getting out of problematic situations nor was the possibility of making errors reduced significantly. This outcome is in line with the tool

specifications [52]. After all, the user is always in charge of making decisions. The system cannot determine if errors are done when solving a task, as it only suggests how to visualize the data appropriately. However, the students reported that the guidance was appropriate for their task and that it helped them reason about the data (UH23–24).

Quantitative Analysis　Following standard practices for developing and assessing surveys and heuristics [26], we performed a quantitative analysis of the results obtained, according to the different quality criteria. In particular, we applied Cronbach's alpha [16] to the results obtained. This is a measure of reliability used to calculate the internal consistency of tests and measures (as our heuristics). We obtained an average score of $alpha = 0.87$, indicating that the collected responses are consistent. This consistency suggests that our heuristics are reliable and that our criteria are not redundant.

6.2.3 Comparative analysis

Comparison with original evaluation　We provide a comparative analysis of the user evaluation results, examining them in contrast to the results of the original evaluation performed by the authors of Voyager [52]. Our evaluation approach provides a much more detailed assessment of the guidance compared to how the system was originally evaluated. The initial evaluation [52] aimed to discover the value of adding guidance and its usefulness in supporting data exploration (i.e., get an overview). The results of the initial evaluation partially cover some aspects, which in our methodology are considered by the criterion of *Relevance*. The authors also assessed the users' trust towards the recommendations (the same aspect is part of our *Explainability* criteria, which has a broader scope, though). In our evaluation, VAUs reported, as in the original evaluation, that they trusted the system recommendations. However, our methodology expands such assessment showing a need for more adaptation, flexibility, suitable explanations, and control over the guidance (as shown in Fig.6).

Comparison with VAEs assessment　We also ran a parallel expert evaluation of Voyager, using the VAEs heuristics and involving three additional visualization and guidance experts. The results are shown in Figure 7. In summary, by comparing results with Figure 6, we can see how the results are consistent between the two groups of evaluators. However, we can also see how the results of some criteria slightly diverge between VAUs and VAEs on some aspects which we describe in the following. Both groups of evaluators reported, for instance, that the guidance was *Timely*, and both assigned low ratings to its *Controllability* and *Explainability*. However, VAUs assigned low ratings to the guidance *Adapatability*, while VAEs, on average, provided higher ratings, signaling that the design of the guidance could be potentially improved. Still, on the *Adapatability* criterion, only one VAEs (1 of 3) recognized that the content of the suggestions does not consider analytic provenance or user preferences, rating it 2. On the other hand, VAUs rated positively the *Relevance* of the guidance in supporting dataset exploration and getting an overview. At the same time, VAEs consistently highlighted that the system could not recognize and prevent errors or biases.

	Explainable	Controllable	Relevant	Visible	Adaptive	Expressive	Flexible	Timely
	1,00	2,00	2,67	4,00	2,00	5,33	7,00	7,00
	3,50	3,20	4,00	4,00	7,00	6,33	7,00	7,00
	5,50	5,20	5,00	6,00	7,00	7,00	6,00	7,00
AVG	3,33	3,47	3,89	4,67	5,33	6,22	6,67	7,00

Figure 7: Results of the expert evaluation of Voyager [52]. Reddish cells represent ratings close to 1 (heuristics violated), while greenish cells represent ratings close to 7 (heuristics applied). Columns are ordered based on the average rating, from lower to higher.

7 DISCUSSION AND LESSONS LEARNED

Importance of quality criteria　The results show that the effectiveness of guidance comes from a combination of different characteristics, which we match to our eight quality criteria. The majority of the experts we consulted characterized the effectiveness of guidance in terms of its

ability to support the completion of tasks (using terms such as *meaningful*, *facilitating*, *goal-oriented*), its transparency to the user (using terms as *clear*, *transparent*, *descriptive*) and its adaptability to the analysis context (with terms such as *user-/task-dependant*, *context-sensitive*). These characteristics can be expressed in terms of the *Relevance*, the *Explainability*, and the *Adaptiveness* of guidance content. Whereas we rated and assigned the same weight to all the criteria, these results suggest that not all of them have the same impact on the final guidance design, and depending on the context some could be prioritized over others. These results provide valuable insights on how users perceive the guidance qualities and therefore support this kind of strategic decision.

VAEs must be really experts　During the design of the evaluation framework, we ran internal tests before running the real evaluations. We encountered some problems during the expert trials, mainly accounting for the lack of knowledge of the guidance concepts. In particular, as mentioned, we phrased the VAEs heuristics using guidance terminology used commonly in the literature. During the tests, we noticed that the two VAEs involved were not entirely familiar with the terminology, which led them to misinterpret some of the heuristics. Although it seems obvious and expected that expert heuristics should be assessed only by real experts, we want to highlight how easy it is to make mistakes and obtain unreliable results. A simple solution to this problem could be rephrasing the heuristics or providing the necessary information about the terminology used before running the evaluation. This issue also raises a point about the necessity to strengthen the dissemination of guidance literature in our community.

Completeness　We combined a bottom-up (i.e., reviewing the literature) and a top-down (i.e., input from the VA scholars) approach for the generation of the criteria, aiming for complete coverage of desirable guidance properties. However, we acknowledge that more work is potentially needed to assess the exhaustiveness of our list. Valuable input for this process could derive from the application of our heuristics in practical scenarios, beyond the initial assessment described in this paper.

Templates for experts and user testing　Finally, we want to highlight that the main contribution of this paper lies not only in the description and discussion of a methodology to evaluate guidance approaches in VA. In addition, we provide a protocol (Sect. 4.2.1) for their application in practical scenarios, for rating and interpreting results, enrolling participants, as well as templates of questionnaires to be used by experts and users for testing in this Git repository [1].

8 CONCLUSION

This paper identifies the crucial factors contributing to the value and effectiveness of guidance. Similar to the value of visualizations [48], designing and developing effective guidance depends on balancing different system features. We identified eight quality criteria, which we instantiated into two sets of heuristics to evaluate and define effective guidance. We ran two evaluations to assess how VAEs and VAUs used our methodology in practice. For the expert evaluation, we created an ad-hoc system with specific guidance features. The results show that the experts were able to coherently identify the strengths and weaknesses of the proposed guidance, showing how our methodology can effectively contribute to guidance design already from its early development stages. We employed a well-known guidance-enhanced visualization tool (Voyager) for user evaluation, asked some VAUs to use it, and evaluated the provided support. VAUs could rate the guidance features consistently, showing a convergence of their and experts' ratings. We show the value of our methodology by providing an in-depth assessment of the provided suggestions.

REFERENCES

[1] Templates and Protocol. https://gitlab.cvast.tuwien.ac.at/dceneda/heuristics-guidance-effectiveness/. [Online; accessed 31-March-2023]. 6, 9

[2] R. Amar and J. Stasko. A knowledge task-based framework for design and evaluation of information visualizations. In *Proc. IEEE Symp. on Information Visualization*, pp. 143–150. IEEE, 2004. doi: 10.1109/INFVIS.2004.10 2, 4

[3] R. Bergeron, J. Gannon, D. Shecter, F. Tompa, and A. V. Dam. Systems programming languages. In M. Rubtnoff, ed., *Advances in Computers*, vol. 12, pp. 175–284. Elsevier, 1972. doi: 10.1016/S0065-2458(08)60510-0 4

[4] F. Bouali, A. Guettala, and G. Venturini. VizAssist: An interactive user assistant for visual data mining. *The Visual Computer*, 32(11):1447–1463, 2016. doi: 10.1007/s00371-015-1132-9 2

[5] D. Ceneda, N. Andrienko, G. Andrienko, T. Gschwandtner, S. Miksch, N. Piccolotto, T. Schreck, M. Streit, J. Suschnigg, and C. Tominski. Guide me in analysis: A framework for guidance designers. *Computer Graphics Forum*, 39(6):269–288, 2020. doi: 10.1111/cgf.14017 1, 2, 3, 4

[6] D. Ceneda, A. Arleo, T. Gschwandtner, and S. Miksch. Show me your face: Towards an automated method to provide timely guidance in visual analytics. *IEEE Trans. Visualization and Computer Graphics*, 28(12):4570–4581, 2021. doi: 10.1109/TVCG.2021.3094870 4

[7] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski. Characterizing guidance in visual analytics. *IEEE Trans. Visualization and Computer Graphics*, 23(1):111–120, 2017. doi: 10.1109/TVCG.2016.2598468 1, 2, 4, 7, 8

[8] D. Ceneda, T. Gschwandtner, T. May, S. Miksch, M. Streit, and C. Tominski. Guidance or no guidance? A decision tree can help. In *EuroVA: International Workshop on Visual Analytics*, pp. 19–23. Eurographics Digital Library, 2018. doi: 10.2312/eurova.20181107 1

[9] D. Ceneda, T. Gschwandtner, and S. Miksch. A review of guidance approaches in visual data analysis: A multifocal perspective. *Computer Graphics Forum*, 38(3):861–879, 2019. doi: 10.1111/cgf.13730 1, 2, 3

[10] D. Ceneda, T. Gschwandtner, S. Miksch, and C. Tominski. Guided visual exploration of cyclical patterns in time-series. In T. Möller and S. Liu, eds., *Visualization in Data Science (VDS)*. IEEE Computer Society, 2018. 2, 3

[11] C. Chen and M. P. Czerwinski. Empirical evaluation of information visualizations: An introduction. *Int. J. of Human-Computer Studies*, 53(5):631–635, 2000. doi: 10.1006/ijhc.2000.0421 2

[12] C. Chen and Y. Yu. Empirical studies of information visualization: A meta-analysis. *Int. J. of Human-Computer Studies*, 53(5):851–866, 2000. doi: 10.1006/ijhc.2000.0422 2

[13] J. Choo, H. Lee, J. Kihm, and H. Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Proc. IEEE Symp. on Visual Analytics Science and Technology*, pp. 27–34. IEEE, 2010. doi: 10.1109/VAST.2010.5652443 1

[14] C. Collins, N. Andrienko, T. Schreck, J. Yang, J. Choo, U. Engelke, A. Jena, and T. Dwyer. Guidance in the human–machine analytics process. *Visual Informatics*, 2(3):166–180, 2018. doi: 10.1016/j.visinf.2018.09.003 1, 2

[15] K. A. Cook and J. J. Thomas. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Lab (PNNL), Richland, WA (United States), 2005. 1

[16] L. J. Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951. doi: 10.1007/BF02310555 9

[17] C. Forsell. Evaluation in information visualization: Heuristic evaluation. In *Proc. IEEE Symp. on Information Visualization*, pp. 136–142. IEEE, 2012. doi: 10.1109/IV.2012.33 2

[18] C. Forsell and J. Johansson. An heuristic set for evaluation in information visualization. In *Proc. Int. Conf. on Advanced Visual Interfaces*, pp. 199–206, 2010. doi: 10.1145/1842993.1843029 2

[19] S. Gladisch, H. Schumann, and C. Tominski. Navigation recommendations for exploring hierarchical graphs. In *Int. Symp. on Visual Computing*, pp. 36–47. Springer, 2013. doi: 10.1007/978-3-642-41939-3_4 1

[20] D. Gotz, J. Lu, P. Kissa, N. Cao, W. H. Qian, S. X. Liu, and M. X. Zhou. HARVEST: An intelligent visual analytic tool for the masses. In *Proc. of Int. Workshop on Intelligent Visual Interfaces for Text Analysis*, pp. 1–4. ACM, 2010. doi: 10.1145/2002353.2002355 2

[21] D. Gotz and Z. Wen. Behavior-driven visualization recommendation. In *Proc. of ACM Conf. on Intelligent User Interfaces*, pp. 315–324. ACM, 2009. doi: 10.1145/1502650.1502695 1

[22] W. Han and H.-J. Schulz. Providing visual analytics guidance through decision support. *Information Visualization*, 22(2):140–165, 2023. doi: 10.1177/14738716221147289 1, 2

[23] T. He, P. Isenberg, R. Dachselt, and T. Isenberg. BeauVis: A validated scale for measuring the aesthetic pleasure of visual representations. *IEEE Trans. Visualization and Computer Graphics*, 29(1):363–373, 2022. doi: 10.1109/TVCG.2022.3209390 2, 3

[24] J. Heer and D. Boyd. Vizster: Visualizing online social networks. In *Proc. IEEE Symp. on Information Visualization*, pp. 32–39. IEEE, 2005. doi: 10.1109/INFVIS.2005.1532126 1, 2

[25] J. Heer and G. Robertson. Animated transitions in statistical data graphics. *IEEE Trans. Visualization and Computer Graphics*, 13(6):1240–1247, 2007. doi: 10.1109/TVCG.2007.70539 2

[26] T. R. Hinkin, J. B. Tracey, and C. A. Enz. Scale construction: Developing reliable and valid measurement instruments. *Journal of Hospitality & Tourism Research*, 21(1):100–120, 1997. doi: 10.1177/109634809702100108 9

[27] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pp. 3363–3372. ACM, 2011. doi: 10.1145/1978942.1979444 1

[28] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proc. Int. Conf. on Advanced Visual Interfaces*, pp. 547–554. ACM, 2012. doi: 10.1145/2254556.2254659 1

[29] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In *Visual Data Mining*, pp. 76–90. Springer, 2008. doi: 10.1007/978-3-540-71080-6_6 1

[30] J. Krause, A. Perer, and K. Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pp. 5686–5697. ACM, 2016. doi: 10.1145/2858036.2858529 3

[31] G. Krishnamoorthy and P. Brusilovsky. Personalized guidance for example selection in an explanatory visualization system. In *Proc. of E-Learn: World Conf. on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pp. 2122–2127. AACE, 2006. 3

[32] M. Luboschik, C. Maus, H.-J. Schulz, H. Schumann, and A. Uhrmacher. Heterogeneity-based guidance for exploring multiscale data in systems biology. In *Proc. IEEE Symp. on Biological Data Visualization*, pp. 33–40. IEEE, 2012. doi: 10.1109/BioVis.2012.6378590 1

[33] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans. on Graphics (Tog)*, 5(2):110–141, 1986. doi: 10.1145/22949.22950 2

[34] J. Mankoff, A. K. Dey, G. Hsieh, J. Kientz, S. Lederer, and M. Ames. Heuristic evaluation of ambient displays. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pp. 169–176, 2003. doi: 10.1145/642611.642642 2

[35] T. May, A. Bannach, J. Davey, T. Ruppert, and J. Kohlhammer. Guiding feature subset selection with an interactive visualization. In *Proc. IEEE Symp. on Visual Analytics Science and Technology*, pp. 111–120. IEEE, 2011. doi: 10.1109/VAST.2011.6102448 1

[36] S. Miksch and W. Aigner. A matter of time: Applying a data–users–tasks design triangle to visual analytics of time-oriented data. *Computers & Graphics*, 38:286–290, 2014. doi: 10.1016/j.cag.2013.11.002 2

[37] R. Molich and J. Nielsen. Improving a human-computer dialogue. *Communications of the ACM*, 33(3):338–348, 1990. doi: 10.1145/77481.77486 2, 4, 6

[38] J. Nielsen. Heuristic evaluation. *Usability Inspection Mehods*, 1994. doi: 10.5555/189200.189209 2, 4, 6, 7

[39] P. O'Donovan, A. Agarwala, and A. Hertzmann. DesignScape: Design with interactive layout suggestions. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pp. 1221–1224. ACM, 2015. doi: 10.1145/2702123.2702149 3

[40] I. Pérez-Messina, D. Ceneda, M. El-Assady, S. Miksch, and F. Sperrle. A typology of guidance tasks in mixed-initiative visual analytics environments. In *Computer Graphics Forum*, vol. 41, pp. 465–476. Wiley, 2022. doi: 10.1111/cgf.14555 4

[41] C. Plaisant. The challenge of information visualization evaluation. In *Proc. Int. Conf. on Advanced Visual Interfaces*, pp. 109–116, 2004. doi: 10.1145/989863.989880 2

[42] B. Saket, A. Endert, and J. Stasko. Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proc. Workshop on BEyond time and Errors: Novel Evaluation Methods for*

Information Visualization, pp. 133–142, 2016. doi: 10.1145/2993901. 2993903 2

[43] D. L. Scapin and J. C. Bastien. Ergonomic criteria for evaluating the ergonomic quality of interactive systems. *Behaviour & Information Technology*, 16(4-5):220–231, 1997. doi: 10.1080/014492997119806 2, 4

[44] 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, 2006. doi: 10. 1145/1168149.1168158 2

[45] F. Sperrle, D. Ceneda, and M. El-Assady. Lotse: A practical framework for guidance in visual analytics. *IEEE Trans. Visualization and Computer Graphics*, 29(1):1124–1134, 2022. doi: 10.1109/TVCG.2022.3209393 1, 4

[46] F. Sperrle, A. V. Jeitler, J. Bernard, D. A. Keim, and M. El-Assady. Learning and teaching in co-adaptive guidance for mixed-initiative visual analytics. In *EuroVis Workshop on Visual Analytics (EuroVA)*, pp. 61–65, 2020. doi: 10.2312/eurova.20201088 2

[47] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady. explAIner: A visual analytics framework for interactive and explainable machine learning. *IEEE Trans. Visualization and Computer Graphics*, 26(1):1064–1074, 2019. doi: 10.1109/TVCG.2019.2934629 4

[48] J. Stasko. Value-driven evaluation of visualizations. In *Proc. Workshop on BEyond time and Errors: Novel Evaluation Methods for Information Visualization*, pp. 46–53, 2014. doi: 10.1145/2669557.2669579 2, 9

[49] C. Stoiber, D. Ceneda, M. Wagner, V. Schetinger, T. Gschwandtner, M. Streit, S. Miksch, and W. Aigner. Perspectives of visualization onboarding and guidance in VA. *Visual Informatics*, 6(1):68–83, 2022. doi: 10.1016/j.visinf.2022.02.005 2

[50] E. Wall, M. Agnihotri, L. Matzen, K. Divis, M. Haass, A. Endert, and J. Stasko. A heuristic approach to value-driven evaluation of visualizations. *IEEE Trans. Visualization and Computer Graphics*, 25(1):491–500, 2018. doi: 10.1109/TVCG.2018.2865146 2, 7

[51] M. Q. Wang Baldonado, A. Woodruff, and A. Kuchinsky. Guidelines for using multiple views in information visualization. In *Proc. Int. Conf. on Advanced Visual Interfaces*, pp. 110–119, 2000. doi: 10.1145/345513. 345271 2

[52] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Trans. Visualization and Computer Graphics*, 22(1):649–658, 2015. doi: 10.1109/TVCG.2015.2467191 8, 9

[53] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager 2: Augmenting visual analysis with partial view specifications. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pp. 2648–2659. ACM, 2017. doi: 10.1145/ 3025453.3025768 2

[54] T. Zuk and S. Carpendale. Theoretical analysis of uncertainty visualizations. In *Visualization and Data Analysis*, vol. 6060, pp. 66–79. SPIE, 2006. doi: 10.1117/12.643631 2

[55] T. Zuk, L. Schlesier, P. Neumann, M. S. Hancock, and S. Carpendale. Heuristics for information visualization evaluation. In *Proc. Workshop on BEyond time and Errors: Novel Evaluation Methods for Information Visualization*, pp. 1–6, 2006. doi: 10.1145/1168149.1168162 2