PREVis: Perceived Readability Evaluation for Visualizations

Anne-Flore Cabouat, Tingying He, Petra Isenberg, Tobias Isenberg

▶ **To cite this version:**

Anne-Flore Cabouat, Tingying He, Petra Isenberg, Tobias Isenberg. PREVis: Perceived Readability Evaluation for Visualizations. IEEE Transactions on Visualization and Computer Graphics, In press, 10.1109/TVCG.2024.3456318 . hal-04665390

HAL Id: hal-04665390
https://inria.hal.science/hal-04665390v1

Submitted on 31 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

PREVis: Perceived Readability Evaluation for Visualizations

Anne-Flore Cabouat, Tingying He, Petra Isenberg, Tobias Isenberg

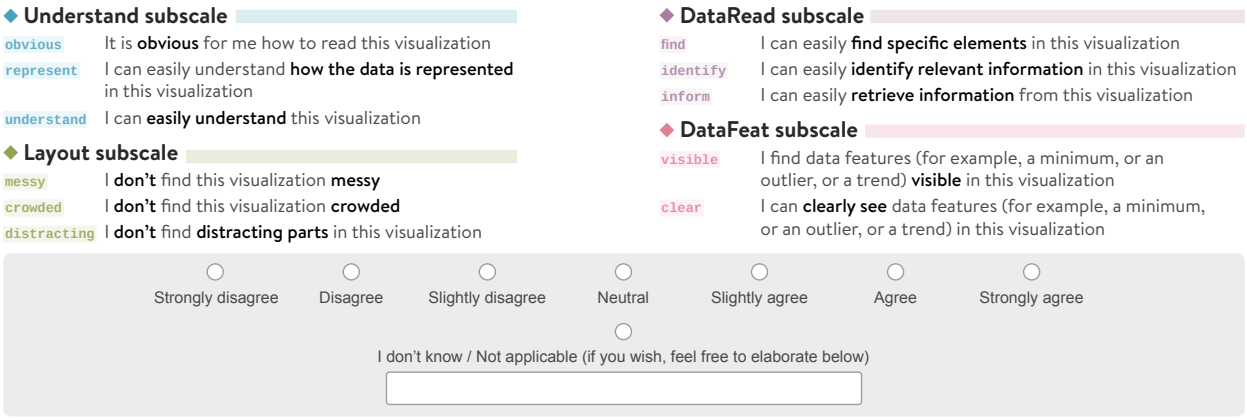

Fig. 1: PREVis subscales, items, and recommended presentation of answer options for a computer-supported questionnaire.

Abstract—We developed and validated an instrument to measure the perceived readability in data visualization: PREVis. Researchers and practitioners can easily use this instrument as part of their evaluations to compare the perceived readability of different visual data representations. Our instrument can complement results from controlled experiments on user task performance or provide additional data during in-depth qualitative work such as design iterations when developing a new technique. Although readability is recognized as an essential quality of data visualizations, so far there has not been a unified definition of the construct in the context of visual representations. As a result, researchers often lack guidance for determining how to ask people to rate their perceived readability of a visualization. To address this issue, we engaged in a rigorous process to develop the first *validated* instrument targeted at the subjective readability of visual data representations. Our final instrument consists of 11 items across 4 dimensions: understandability, layout clarity, readability of data values, and readability of data patterns. We provide the questionnaire as a document with implementation guidelines on osf.io/9cg8j. Beyond this instrument, we contribute a discussion of how researchers have previously assessed visualization readability, and an analysis of the factors underlying perceived readability in visual data representations.

Index Terms—Visualization, readability, validated instrument, perception, user experiments, empirical methods, methodology.

1 Introduction

When looking at examples of data visualizations, it is intuitively clear that some are easier to read than others. For many data analysis use cases, poor readability will drastically reduce the usefulness of a visual representation of data for the viewer. As such, readability is a basic quality criterion in data visualization [53]. One of the fundamental challenges in studying the readability of data visualizations, however, is that the concepts of *reading* and *readability* are held as tacit knowledge. The terms are often used in scientific writing without clear definitions of what they specifically mean in the context of data visualization—recalling Kosara's "empire built on sand" [54].

Readability of text is broadly defined as "the quality of being easy and enjoyable to read" [19]. It applies to letters and words as well as entire books. Linguists have developed hundreds of formulas to analyze the readability of texts [8], but this approach fails to take into account characteristics of readers. Since readability is better explained as a function of the interaction between the properties of texts and the characteristics of readers [5], researchers now seek to analyze text difficulty based on cognitive theories. Such an approach may also be suitable to explore the readability of visual representations of data.

• *Anne-Flore Cabouat, Tingying He (何汀滢), Petra Isenberg, and Tobias Isenberg are with Université Paris-Saclay, CNRS, Inria, LISN, France. E-mail: given_name.family_name@inria.fr*

Manuscript received xx xxx. 202x; accepted xx xxx. 202x. Date of Publication xx xxx. 202x; date of current version xx xxx. 202x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.202x.xxxxxxx

As we discuss in more detail below, a few definitions of "readability" exist in the visualization domain, yet they do not fully overlap. As a result, it is unclear to what extent different approaches to measuring readability can thoroughly capture the concept. In addition, we do not have a definition of what "reading" a data visualization is as a cognitive activity. Cognitive processes in visualization range from low-level visual perception [83] to high-level activities such as data exploration [111], insight and knowledge generation [86, 97], sense-making of unfamiliar visualizations [58], or decision-making [74]. Current cognitive models of visualization comprehension [37, 48, 74] provide important theoretical grounding to explain how people process information from visual data representations; the models, however, do not specify the boundaries of "reading" within the cognition continuum.

Our work is based on the fundamental premise that readability is a crucially important quality criterion in data visualization. As such, it requires formal definition and ***empirically verified*** methods to study it. In this paper we present the development and validation of our PREVis questionnaire. PREVis is a reliable instrument that allows respondents to rate how readable they find a static data visualization across 4 dimensions: layout clarity, ease of understanding, ease of reading data features, and ease of reading data values. During the development process, we also had to take first steps in clarifying what readability means in data visualization. This clarification is important because discrepancies in the use of terminology pose issues of comparability and reliability of empirical findings. In particular, we observed that researchers who asked participants to rate the readability of visualization did so using a wide variety of terms and answer options. Our PREVis tool addresses this problem because we followed well-established methodologies in scale development [12, 30]. Developing a valid scale

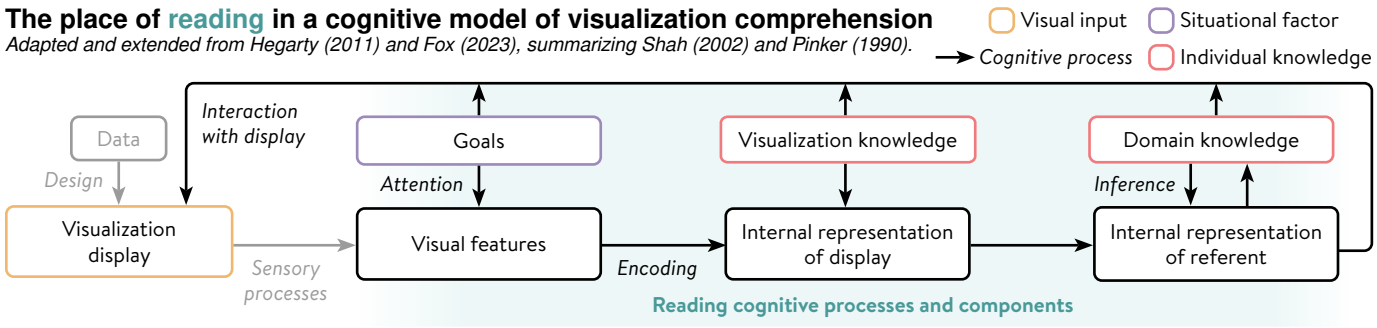

Fig. 2: Our proposition to place *reading* processes within the model proposed by Hegarty [48], summarizing Shah *et al.* [89] and Pinker [78].

requires researchers to review existing literature and collect experts' suggestions on how to phrase the scale's items. In that way, it has allowed us to better frame the concept of readability in data visualization. While we do not promise to define readability entirely—yet—, we posit that it is not only a characteristic of the visualization; it is also highly individual and situational. As such, in our work, we talk about the concept of *perceived readability*, and we emphasize that measures of readability should take the reader's abilities and goals into account, in addition to measurable characteristics of the visualization display—for example, the density of dots in a scatterplot, or the number of edge crossings in a node-link graph.

In summary, with our work we make the following contributions:

- a validated instrument to study perceived visualization readability that can be deployed in empirical research on data visualization,
- an analysis of the factors underlying perceived readability of visual data representations,
- an in-depth discussion and analysis of the concept of readability and how it has been used in the visualization field, and
- an example of careful application of statistical methods from the psychology research field, and how it can contribute to advance theoretical models of data visualization.

2 Related Work

There are three major related fields that are important to our work: definitions of readability, ways of measuring readability, and scale development. We review these fields next.

2.1 Defining readability

Readability broadly refers to "the ease of reading" in the context of written words. Reading is a complex behavior [33], making it difficult to measure and research. Reading theorists describe how readers decode letters and words as well as how they integrate them into text comprehension [76]. Understanding a text further affords the reader with the ability to produce inferences and complex reasoning. While *legibility* essentially focuses on the ease of distinguishing letters, readability can relate to individual words as well as entire books.

In the context of data visualization, researchers have proposed several models of information processing for graph comprehension [38, 55, 78]. In these publications, the authors commonly refer to the viewer of a visualization as a "(graph) reader" but do not define the boundaries of the reading behavior itself. Conversely, Curcio [28] explicitly refers to three levels of reading—reading the data, reading between the data, and reading beyond the data—but does not propose a cognitive model of the processes involved. Drawing from the literature on reading texts, we posit that the *reading* of visual representations encompasses all processes that allow the reader to transform the visual features retrieved from early visual processes into a meaningful internal representation of the information displayed in the visualization [18]. In Fig. 2 we propose a first attempt at situating reading processes in a visual display comprehension model [48].

Yet readability still lacks a formal definition in the data visualization context. Next, we thus review the different components that influence the ease of reading visual representations [37]: the display, the individual, and the task. We then review existing definitions of readability.

2.1.1 The influence of display on readability

Visualization researchers view readability primarily as a quality of the visual object, and have extensively studied the effectiveness of visual encodings for data visualization since early work on the classification of visual variables [10, 24, 64]. Recent perception studies suggest a need to refine these important foundations because visual variables may interact with each other [94] and with other cognitive processes such as attention [47]. Another related challenge in visualization design is to avoid the delivery of overwhelming amounts of information by managing data complexity [14]. For example, Henry et al. [49] refer to readability as *visual complexity* and propose a solution to reduce this complexity in social network representations. To account for the influence of display on readability, in our work we tested our scale items on visualizations showing different amounts of data points and visual variables during the scale development phase.

2.1.2 The influence of individuals on readability

Reading is not just seeing: it involves semantic understanding of the content to construct a coherent mental representation [38, 48]. Accordingly, readers need domain knowledge about what is represented—the *referent*—and skills to build an internal representation of the display. The ability to construct and use mental representations across various types of data visualizations is called *visualization literacy* [13], or *graphicacy* in psychology [81]. A growing stream of research focuses on testing and understanding factors of visualization literacy [6, 36, 59]. Yet many questions remain open regarding this skill [95], including how visualization literacy interacts with other literacies (e. g., numeracy) and how it affects the reader's judgment of a data visualization's trustworthiness. Several individual traits may also play a role in reading data visualizations [62, 75], such as spatial abilities [45, 102], or verbal working memory [98]—a measure of the ability to store and manipulate verbal information. While we do not know yet how these factors integrate with existing models of visualization comprehension, we tested our candidate scale items on common and less common visualizations.

2.1.3 The influence of task on readability

Finally, a reader's goals also have an impact on how they approach a visual representation. There is always a reason why people engage with data visualizations [15]: it can be as diverse as the need to acquire knowledge, the desire to communicate ideas to others, or sheer aesthetic pleasure. Researchers have long noticed that the task at hand interacts with visual encodings when a reader processes information from a visualization [92], and can influence perceptual processes [39]. Recently, Quadri and Rosen [83] surveyed 132 perception studies and distributed them across 11 low-level tasks from Amar *et al.* [3]. In a study using visualizations from the MASSVIS data set, Polatsek *et al.* [79] found that readers' eye gaze patterns were more closely related to the task at hand than to the visual saliency of objects in the visualization. Wang *et al.* [106] conducted a systematic study on scatterplots and noticed that the influence of the visual design and the dataset's size on participants' accuracy and response time were different depending on the comprehension task at hand. In line with existing knowledge on text readability [68], good readability of a visualization will entail different design requirements depending on the reader's goal. Since readability is thus dependent on tasks, in our work we ensured during the scale

development and validation tests that participants perform reading tasks before they rated the perceived readability of a visualization.

2.1.4 Definitions of readability

Despite all of this past work, only few formal definitions of the concept of readability in visualization exist. We are aware of the following:

- *"The relative ease with which the user finds the information he is looking for"* by Ghoniem *et al.* [40], who proposed this definition in the context of comparing user reading performance for matrices with node-link representations of graphs. Similarly, Tu and Shen [103] based a definition on eye movement in their work on treemaps: *"how easy to visually scan a layout to find a particular item, based on how many times viewers' eyes have to change scan direction when traversing a layout."* Although both definitions emphasize the central role of the reader's goal, they focus on the visual query and do not include the ability to make sense of retrieved visual objects.
- *"The ability to make direct observations from the visualizations"* [85]. Ruchikachorn and Mueller proposed this definition to characterize reading tasks in their work on the use of analogies for teaching novices how to read unfamiliar visualizations. Its minimalist phrasing applies to a wide variety of visual representations, and with the word "direct" they seem to remove interactive features from their scope. It describes, however, a learner's ability rather than a property of the interaction between the reader and the visual object. In that sense, it appears to be closer to a definition of visualization literacy coupled with domain knowledge than to a definition of readability.
- *"The extent to which a visualization supports the graphical perception of the information it contains"* [101]. In the context of assessing readability of stacked steamgraphs, Thudt *et al.* proposed this definition centered on the visual object, and added the important notion of *information*. Yet, this definition fails to capture the individual nature of readability as there is no explicit mention of a reader. We also note that this definition relates to a previous definition of *graphical perception*: *"the visual decoding of information encoded on graphs"* [24], which does not encompass further understanding.

While none of these definitions fully captures the *perceived readability* construct we study in this work, each contributes to its broad scope.

2.2 Measuring readability

User studies in data visualization research frequently involve measuring the readability of visual designs. The most commonly reported considerations are participant task performance, layout metrics, subjective feedback, and, to a lesser extent, eye-tracking recordings.

Task performance assessment consists of measuring task completion time and answer accuracy for data visualization tasks—such as retrieving a value, detecting a trend, or comparing two visual components in the visualization. For example, Bu *et al.* [16] assessed the readability of stacked area graphs based on accuracy and completion time for three tasks designed by Thudt *et al.* [101]: read the thickness of an individual layer, read the variation of an individual layer's thickness, and read the overall thickness of aggregated layers. A few studies evaluated readability based on accuracy alone [93], but more often researchers collected and analyzed both time and accuracy (e. g., [16, 49, 105, 107]).

There are some limitations to using task performance as a proxy for readability: the usability of interactive features can arguably affect completion speed, a participant's prior beliefs can influence the accuracy of their response [110], and the task at hand might extend beyond the scope of reading (i. e., mental calculations). To put it briefly, task performance on a data visualization may be *influenced by*—but is not a *targeted measure of*—readability, as many co-factors play a role.

Layout drawing metrics are also used to assess readability by approximating representations' desired visual properties. In node-link visualizations (e. g., [4, 44]), these are called aesthetics metrics. They include edge crossings, node overlapping, neighborhood preservation, or global symmetry [9, 82]. Beyond node-link layouts, Giovannangeli *et al.* [41] proposed a drawing algorithm to improve the readability of scatterplots, expanding from previous work on overplotting reduction [60], and Goffin *et al.* [42] computed metrics to quantify how different placements of word-scale visualizations affected the readability of documents. Still, the range of such metrics is necessarily limited, and existing evaluation instruments cannot be applied to novel types of representations. This estimative approach also does not consider individual factors influencing readers' visual perception and understanding.

Eye-tracking recordings provide spatio-temporal and physiological data that allow researchers to explore the visual behavior of visualization readers [52, 56], and to evaluate the usefulness of visual representations [66]. This technique is already an established method in cognitive science to study reading strategies on visual displays [65, 72]. Eye-tracking studies contribute to building datasets, which can then be used to test cognitive saliency models [63, 79] or to train AI models for human gaze prediction [91]. However, this method requires heavy logistics both for conducting a user study and analyzing the resulting data, and the eye-tracking device can feel invasive for the participant.

Finally, visualization researchers commonly use **subjective assessments** in experiments [51, 57] as such data contributes to building a more holistic understanding of how people engage with data visualizations [87]. Subjective assessment methods include sketching observations [21], semi-structured interviews [109], think-aloud protocols [88], and open-ended comments in surveys [1], often associated with rating questions. During the preliminary step of our present work, we reviewed 34 studies with at least one question aiming at collecting subjective ratings relevant to the concept of readability (as shown in Table 4 in Appx. B). The questionnaires we retrieved not only varied greatly in terms of wording or number of questions asked; we also found a diversity of answer options and numbers of rating categories. For example, some researchers asked people to agree on several statements using a 7-point Likert scale from "Totally Disagree (1)" to "Totally Agree (7)" [16]; others asked people to rate visualization using 5 unnamed points between polar opposite terms such as "Well-organized" and "Poorly-organized" [73]; and even others asked participants to rank 3 visualizations from 1 to 3 based on their "legibility," with the possibility to give them equal rankings [77]. Such discrepancies prevent a comparison of results across studies. Standardized and validated tools to measure readability of data visualizations from the reader's perspective solve this problem—which is what we aim to provide with the validated measuring instrument we develop in this work.

2.3 Scale development

Researchers use psychological scales to measure a specific, identified construct among the respondents (the "latent variable") [30]. Scales involve a set of items, each item typically consisting of a statement or a question with matching Likert scale response options. Scales can increase statistical power, granting reliable results for smaller sample sizes, which is a particularly relevant issue for usability evaluation studies [2]. As scales provide standardized results, they also facilitate the replicability of studies and comparison of results in a meta-analysis. But a scale can only provide the aforementioned benefits if researchers follow a rigorous approach to develop and validate its content and its form. To ensure a high validity and reliability of the final instrument, we rely on recent scale development work in our field [46] and follow best practices in scale development methodology [12, 30] at every step.

3 What is PREVis and how to use it

Following established methodologies for scale development [12, 30], we contribute PREVis, a validated instrument for comparing the perceived readability of different visualizations. As we show in Fig. 1, it comprises 11 items across four subscales that cover the following concepts (i. e., "dimensions": see our glossary in Appx. A):

- **Understand:** the intelligibility of the encodings for the reader
- **Layout:** the visual clarity of the layout
- **DataRead:** how easily people feel they can read data values
- **DataFeat:** how easily people feel they can read data patterns

PREVis can be used in experiments, after participants completed reading tasks, to measure their perception of a data visualization's readability in a standardized way. It can also complement qualitative work to evaluate user experience when developing a new technique. PREVis is not meant to replace existing empirical tools in data visualization

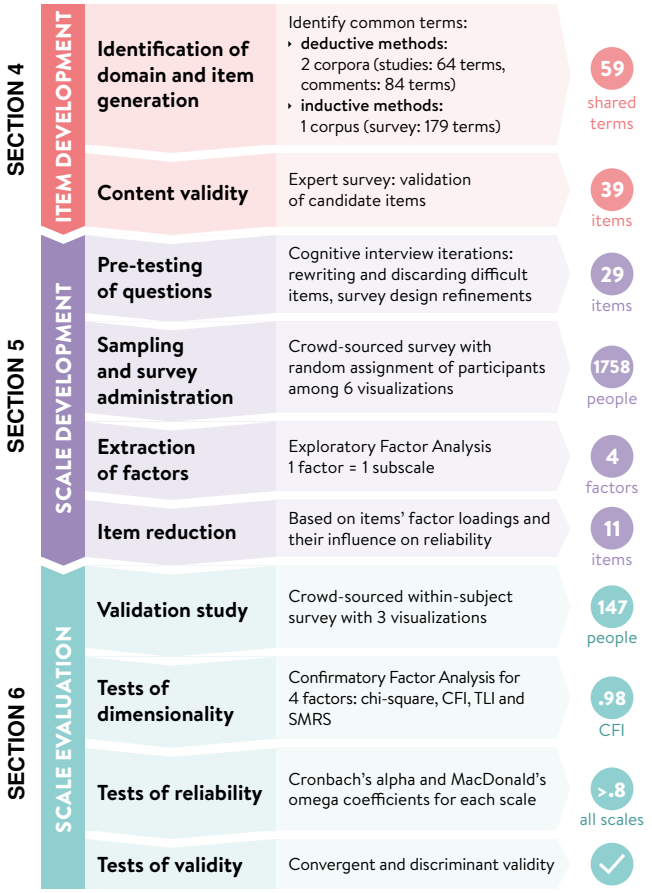

Fig. 3: Summary of our method (adapted from Boateng *et al.* [12]).

research. Instead, our instrument can be used to complement other measuring approaches such as task performance or user interviews and help researchers build a more holistic view of their empirical findings.

Together, the four concepts in PREVis capture perceived readability more holistically, but researchers can also use each subscale independently to evaluate the corresponding concept. Each PREVis item should be presented using a fully-labeled 7-point Likert scale. We recommend to include an "I don't know" option with a text field to collect qualitative feedback when participants feel they cannot answer, and to treat such answers as missing data (N/A) for quantitative analysis purposes. For each subscale, researchers can obtain a single score by averaging individual items' ratings. This single score should be viewed comparatively and not interpreted as an absolute measure of the related concept. We do not recommend to calculate an average score between scales as these capture separate dimensions of readability.

To create and validate PREVis, we followed a three-phase, nine-step process outlined in Fig. 3. We introduce each step in detail in the following sections. We provide details of our methodology and results in our OSF Research log, definitions for the technical terms we use in Appx. A, as well as detailed usage guidelines at `osf.io/9cg8j`.

4 Item Development

In this first phase of the process, we established the boundaries of the readability concept, identified 59 relevant terms and generated a set of 39 candidate items. To start, we sought to delineate the domain of our scale. We collected definitions and models of *reading* and *readability* from the dictionary, the psychology literature, and the data visualization literature. Beyond the summary of our findings from Sect. 2.1, we offer more details on our explorations in the *Domain definition* section of our OSF Research log. In the absence of an existing theoretical framework for our question, methodological guidelines on scale development recommend that the domain be specified *a posteriori* [12]. Accordingly, we moved on to the next step: to identify a pool of suitable terms.

4.1 Relevant terms identification

To generate a list of rating items relevant to assess readability in data visualization, we first needed to establish a pool of relevant words to use. There are two general ways to identify appropriate terms for scale items: deductive and inductive methods [12]. We used both and constitute 3 different pools of terms, as we show in Table 1. We provide details and data files in our OSF supplemental material folder.

For each corpus, we followed a similar selection process: the first author began by reviewing the corpus' content and retrieving broadly relevant keywords or expressions. All authors then collectively reviewed this work and selected words according to the inclusion criteria described further below. We refer to the selected words as the collected *terms*. At this stage, we intentionally used over-inclusive criteria because the scale development methodology later allows us to detect weakly related items and it is important not to miss any variables that should have been included [23]. After each item collection, we applied *stemming* [80] to reduce the different variations of a word to its root. For example, the words "distracting" and "distracted" both yield the "distract" root (called *stem*). We refer to the unique stems as *unique* terms in the following paragraphs for the sake of reading ease.

4.1.1 Deductive method

We conducted a literature review on IEEE VIS papers (1990–2022) and TVCG and CG&A journal papers presented at IEEE VIS (2011–2021). With the Boolean query `"likert" AND ("readab" OR "legib" OR "deciph")` we found 128 papers, from which we derived two sets of terms: terms used in study questionnaires and terms from reported comments of participants. 13 publications provided words for both pools of terms. We included terms from questions and comments related to the easiness of reading, perception, and understanding, as well as the effectiveness or efficiency of visual elements for a given task. We excluded questions and comments related to the visualization system (e.g., interactivity, general preference) or to aesthetic judgment.

Terms from study questionnaires (Pool 1). For this pool we focused on collecting words used in items of rating questions in user studies. In addition to the inclusion criteria above, we included terms from questions that authors related to readability in the method, results, or discussion sections. We expected the term "visualization" to be over-represented and we thus excluded it (see Appx. B for more details on the process and results). We collected 135 terms (64 unique ones) from 34 publications (Table 4 in Appx. B).

Terms from participant comments (Pool 2). For this pool, we focused on reported comments from participants. Texts in this corpus were sometimes direct quotations from participants comments, and sometimes we reworded them to summarize comments from multiple participants. Beyond the exclusion criteria we listed above, we also excluded terms such as "usable" as it would be difficult to assess whether it was related to the visual representation alone or to the whole visualization system, including interactive features. We collected 165 terms (84 unique) from 34 publications (Table 5 in Appx. B).

4.1.2 Inductive method

To complement our literature review with input from experts, we conducted a survey, which we pre-registered (`osf.io/4dcav`) and for which we received IRB approval (Inria COERLE, avis № 2023-17).

Participants. We invited 106 visualization experts by direct e-mail to participate in our survey. We selected them based on our knowledge of their work and their reputation in the visualization community, ensuring that their expertise covers a wide range of topics. We did not compensate participants for taking part in the study. After sending the invitations, we waited for 10 days and, during this time, received 29 complete responses (experience in the field: mean 17.5 years; 6 women (cis or trans), 21 men (cis or trans), 2 rather not answer). All responses were valid and we included them in our analysis.

Procedure. We first asked participants to complete the informed consent form and to answer background questions about their gender and years of experience. We then explained the study scenario, which involved wanting to study people's perception about the readability of a visualization they created, using a 7-point Likert scale with the

Table 1: Summary of our terms collection.

method	corpus	sources	terms (unique)
deductive *literature review: 128 publications*	questionnaires in studies (Pool 1)	34 studies	135 (64)
	reported participants' comments (Pool 2)	34 studies	165 (84)
inductive *expert survey: 132 statements*	collected propositions of items (Pool 3)	29 experts	447 (179)
Sum of collected terms (overall unique terms)			**747 (249)**

question: "To what extent do you agree or disagree with the following statement: [...]." We asked each participant to give us at least three statements they would use to fill in the blank in the question. We also gave them the opportunity to leave additional comments after providing us with their items suggestions, if they chose to do so.

Results. From the 29 completed surveys we collected 132 items and 13 additional comments. We found 3 items aimed at assessing whether the visual object is a data visualization or not; as this was not in the scope of our work, we discarded these 3 items. Then we split the remaining 128 items into 147 statements (e. g., "I can process the data elements quickly to form an overview of the result." was separated in two statements: "I can process the data elements quickly" and "(I can) form an overview of the result"). We used this corpus of statements to establish our third pool of relevant terms, and to find conceptual and phrasing patterns that would inform the writing of our scale's items.

Terms from expert survey (Pool 3). For this pool, we worked on the assumption that experts proposed content highly relevant to our work. Thus, we only had a few exclusion criteria: we did not include terms describing the visualization (e. g., "picture", "chart"), common verbs (e. g., "can" or "have"), and terms from secondary prepositional phrases (e. g., in the statement "When I look at this image, I immediately understand how to recover the data values" we did not retrieve any word from the phrase "when looking at the visualization"). We extracted 381 keywords or key expressions (e. g., from the previous example we extracted "immediately understand; how to; recover; data values"). With this process we collected 447 terms (176 unique terms). We provide this table as a separate file in our supplemental material, as it is too long for comfortable reading in a paginated document.

4.2 Item generation

We aligned all unique terms from our 3 pools and filtered out terms present in only one pool. We obtained a list of 59 unique terms common to at least 2 pools (Table 6 in Appx. B). We excluded 12 irrelevant or overly redundant terms (e. g., we excluded "hard" as it was too redundant with "difficult"). From the remaining 47 terms, we identified 35 primary terms (e. g., "clear", "understand") and 12 auxiliary ones (e. g., "data", "easy") as we describe in our OSF Research log.

Phrasing and conceptual patterns in expert propositions. Creating scale items does not only require a set of relevant terms; researchers also need to choose how to combine the terms and write sentences that can be used as appropriate rating items [30]. To inform such choices in the generation of our candidate items, we manually analyzed how experts phrased their statements through syntactic roles and conceptual families of words. We provide an example of this work in Appx. C, and a more detailed description in our OSF Research log. Based on this process, we obtained a summary of the experts' statements in the form of hierarchical data representing the flow of sentences, aggregated by conceptual families of terms. We used this data to help us answer questions such as:

- What kind of objects are referred to as being "clear"?
- Are there frequent descriptions of how people should be able to "understand" readable visualizations (e. g., quickly)?
- Should we write one or two items with the term "inform", which stems from two words: "information" and "informative"?
- Should we write "This visualization is understandable" or "I can understand this visualization"?

Referring both to the statements summary and to the original sources of terms, the first author generated a first draft of possible items, with at least one possible item being based on each of the 35 primary terms. When relevant according to the statements summary, we also included items based on auxiliary terms. The team then discussed all items and selected 39. In particular, we discussed the drawbacks of using reverted items such as "I find this visualization complex to read." We decided against rewriting these items to keep them as easy to read as possible until the pre-testing user study (see Sect. 5.2). Finally, we harmonized the wording of all 39 items as we show in Table 7 in Appx. C.

4.3 Item validation

Experts are highly knowledgeable in the domain, and are also potential users of our final measuring instrument. It is thus recommended [12] to seek validation of the generated items among expert judges. To that end, we conducted a study, again pre-registered (`osf.io/d9nmu`) and IRB-approved (Inria COERLE, avis № 2023-17).

Participants. We invited the same 106 visualization experts as before to participate in our survey by direct e-mail. Again, we did not compensate participants for taking part in the study. After sending the invitations, we waited for 30 days and, during this time, received 31 complete responses (experience in the field: mean 18.9 years; 8 women (cis or trans), 20 men (cis or trans), 1 non-binary, 2 rather not answer).

Procedure. We first asked participants to agree to a consent form and to provide background information about their gender and experience in data visualization. We then displayed the list of 39 items and asked them to rate each item according to its relevance for describing the perceived readability of a data visualization, using a 1–5 point Likert scale (1 = "not at all relevant," 5 = "very relevant"). Participants could leave additional comments after providing us with their ratings. A few comments pointed out that the item "I find parts of the visualization distracting" should have been marked as a reverted item, which was an oversight on our part when finalizing the survey.

Results. We calculated the mean, the mode, and the median score for the 39 items. All means were above 3, and all modes and medians were above or equal to 3 (see Table 8 in Appx. D). As a result, we kept all 39 items for the next stage of scale development.

5 Instrument development: exploratory phase

When creating a scale, the goal of this phase is to establish the underlying dimensions of the construct—in our case, "perceived readability"—and to select the most appropriate items to measure it. As we found from our related work readings, multiple factors are likely to influence the perception of a visualization's readability, e. g., the clarity of a visual design, or the reader's ability to understand how the data is represented and its topic, or the difficulty of the reading task at hand. To make this issue even more delicate, such factors may also be correlated among themselves: a reader may, for example, feel that a visual design is unclear because they are not familiar with the type of representation, or they might find it difficult to observe data patterns because of visual clutter. In addition, 39 items would be too many for the easy-to-administer measuring instrument we envisioned. We thus needed to reduce our pool and to select items that can best reflect how readable respondents find a given data visualization. Exploratory Factor Analysis (EFA) [108] was specifically developed to help researchers determine the underlying factors in a measured construct and to help them identify the most relevant variables—in our case, scale items—to measure it. In scale development, measures of reliability such as Cronbach's alpha (α) and McDonald's omega (ω) complement the EFA approach to select items which will best contribute to the instrument's reliability [30]. To collect data for conducting such types of analysis we ran a third experiment, again pre-registered (`osf.io/4dcav`) and IRB-approved (Inria COERLE, avis №2023-17).

5.1 Experiment design

The goal of this study was to collect participants' ratings of perceived readability for several data visualizations for conducting EFA and reliability analyses. As our goal was to develop an instrument that could be used to measure perceived readability (1) across multiple

populations, (2) with diverse visualization idioms, and (3) across a wide range of readability levels, we now describe our choices regarding these 3 elements in our study design.

Target population. Candidate scale items should be tested on a heterogeneous sample of the target population [12]. We focused on the general population as a baseline because we wanted our instrument to be useful across multiple populations in research. In particular, we decided not to apply any exclusion criteria other than English language fluency. We decided to recruit participants from Prolific and to conduct our study online, as we describe in more detail in Sect. 5.3.

Visualization stimuli. Here, we refer to the visualizations we asked participants to rate with candidate scale items as "stimuli." As our envisioned instrument should capture information about perceived readability across a broad spectrum of data visualizations, we needed to test the items among various stimuli—i. e., across visualizations presenting variability in aspects that are likely to impact readability. In the absence of an existing objective instrument to evaluate readability for a diverse set of visualizations, we focused on variations in the underlying data (number of data entities and attributes represented) and on the visual encoding (number and appropriateness of visual variables used to encode the data attributes and expected familiarity of visualization idiom). As a result, we used the 6 visualizations in Fig. 1 as stimuli for this study; for their characteristics and our design rationale see Appx. F.

Reading tasks. We needed to ensure that participants would at least attempt to read the data visualization, before asking them to rate their perception of its readability. For an online survey, it meant that we would have to give them reading tasks to perform, before showing the rating items. As this reading experience would shape their opinion on the readability of the visualization, we needed to ascertain that we would only use tasks that are within the scope of "reading." For instance, a task that requires additional mental calculation such as evaluating an average or a sum would not be appropriate. We reviewed three main taxonomies of visualization tasks [3, 15, 28] to identify the following list of possible reading tasks: *retrieve value*, *find extremum*, *determine range*, *characterize distribution*, *find anomalies*, *cluster*, *find trend or correlation*, and *make comparisons*. As an additional criterion, we wanted the reading tasks to be relatively simple and quick to complete for participants, regardless of their experience or skill in reading visualizations. The quality of the collected ratings would be crucial at this stage of our work, and careful reading and answering of 39 rating questions would require sustained attention from the participants. As fatigue has been documented to appear after 10 minutes in crowdsourced studies [113], our goal was to allow respondents to complete the survey under this threshold. While acknowledging that the difficulty of the reading task at hand might have an impact on how readable participants find a visualization, our main concern was to allow respondents to stay focused throughout the entire survey. Easiness of a visualization task is not absolute; instead, it relates to how appropriate a visualization is for a task. As such, we referred to the work from Lee *et al.* [59] on building a Visualization Literacy Assessment Test (VLAT), to select two easy reading tasks for each stimulus, as we describe in more detail in Appx. F. To sum up, we retained two criteria when designing our tasks: easiness and adequacy to the scope of reading.

5.2 Pre-testing

Pre-testing items before administrating a survey is a crucially important step: it helps to ensure that items are actually meaningful to the target population [12]. The goal of pre-testing is to revise the phrasing of items to maximise their clarity, eliminate items that cannot be improved, and examine the extent to which people are able to use the answer options to produce ratings. As such, it is also a way to integrate insights from members of the target population in the scale development process. For this study we recruited 11 participants and conducted cognitive interviews [7], a form of think-aloud protocol dedicated to evaluating questionnaires. As a result, we made changes to the items in-between rounds of interviews. In particular, we reworded reverted items to negative phrasing, we dropped 11 items, and we created two items to replace the "cluttered" term with two related and clearer terms: "crowded" and "messy." Participants' feedback also allowed us to refine the presentation of stimuli, the survey's user interface, and the Likert-scale options. For reasons of space, we include the details of this study's design, procedure, and outcomes in Appx. E. The study received approval from our IRB (Inria COERLE, avis №2023-17).

5.3 Survey administration

From the pre-test study we obtained a final set of 29 items, which we used to conduct our exploratory survey. We ran the survey in two separate rounds with 6 stimuli described in Sect. 5.1 and Appx. F.

Participants. It is difficult to find a consensus on how large a sample should be for an exploratory study. A general rule is that the more items one wants to test, the more participants are required. In line with suggestions from our methodological references [12, 30], we targeted a sample size of 300 participants per visualization. We recruited participants from Prolific, who had to be fluent English speakers and of legal age. Participants received a compensation of €11.52 per hour.

Procedure. After answering a consent form and a question about color-vision deficiency, each participant was randomly assigned to one of the 6 possible stimuli. A short contextual description and a title complemented each stimulus image. We asked participants to answer 2 reading questions and 1 comprehension check question about the visualization. Regardless of their answers to the reading questions, participants had to answer the comprehension check correctly within two attempts to be able to move on to the next part of the survey. In that final section of the survey, we asked participants to rate the visualization using our 29 candidate items, randomized with one attention check item. Participants answered on a 7-point agreement scale. Each point was labeled, from "strongly disagree" to "strongly agree," and there was a separate option labeled "I don't know / Not applicable" with a short text field. We share additional details in Appx. F (e. g., a screenshot in Fig. 13), and printouts of the surveys in our supplemental materials.

5.4 Survey results

We recruited a total of 1,801 participants, who all provided their informed consent. Due to inconsistencies between Prolific and our collected data, we removed data for 10 participants in the first deployment round. In addition, we excluded 33 participants from the analysis who failed attention check questions. As a result, we included ratings from 1,758 participants (ages: mean 32.2 years, SD 10.9 years; 39% female, 59% male, 2% non-binary, and <1% gender not disclosed; education: <1% no formal education, 5% secondary education, 20% high school diploma, 10% technical or community college, 41% undergraduate degree, 22% graduate degree, 2% doctorate); color-vision deficiency: 3% yes). Due to our random assignment of participants to our 6 stimuli, each stimulus received 293 valid ratings on average (SD = 8.27).

Missing data. As we offered an option to answer "I don't know" to rating questions in the exploratory survey (Fig. 13 in Appx. G), there was missing data in our collected ratings. As pre-registered, we followed guidelines from Mirzaei *et al.* [69] on handling such cases by calculating the amount of missing data and testing if it was "Missing Completely at Random" (MCAR) using the `misty` package in `R`. We did so survey-wise and stimulus-wise. Missing data was not MCAR for three stimuli, but it was always negligible (< 1%). We report the details of this analysis and subsequent missing data treatment in Appx. G.

5.5 Instrument dimensions and item reduction analyses

Item reduction analysis in scale development aims at maximizing the instrument's measuring accuracy—i. e., minimizing the measurement error, while minimizing its length—and, thus, the time required for respondents to answer all items. Two theories provide tools to assist scale development [12]: Classical Test Theory (CTT) and Item Response Theory (IRT). Each theory's framework provides a different, complementary approach for assessing the measuring performance of items, but they both share the fundamental requirement that scales are *unidimensional* instruments [30]. In other words, if items are to be combined into a scale, they must reflect one—and only one—construct. When the scale's domain is defined in pre-existing theoretical work, researchers can integrate the fundamental unidimensionality pre-requisite from the start of the item development phase. As a result, a common

Table 2: Survey-wise and stimulus-wise parallel analyses suggest that 3 to 5 factors are necessary to explain the data better than at random.

Full survey	A	B	C	D	E	F
5	5	4	3	4	4	3

practice for the scale development phase consists of first analyzing item correlations to reduce the pool of items, before conducting tests of dimensionality [12]. We, however, did not have a theoretical framework defining readability in data visualization and, based on our preliminary investigations exposed in Sect. 2.1, we had suspicions that "perceived readability" might not be a unidimensional construct. Therefore, we first checked that all items were loosely correlated (above 0.3, as shown in Fig. 14 in Appx. H), followed by conducting EFA on the collected data. Only then did we proceed with item reduction analysis, based on factor loadings and reliability tests for each individual factor.

5.5.1 Exploratory Factor Analysis (EFA)

We conducted our main analysis on the full dataset from valid participations in our survey. As an additional exploratory analysis, however, we also conducted EFA on each individual stimulus' dataset. This extra precaution allowed us to confirm that statistical analyses converged towards 3–5 factors. Finally, we conducted a multi-group Confirmatory Factor Analysis (CFA) to determine whether our questionnaire elicited similar response patterns across stimuli, thus serving as a confirmation of our factor structure before we proceeded with the final item reduction analyses. We followed Watkin's best practices guide [108] for conducting EFA. We provide the `R` and `Python` notebooks we used in this process as a part of our supplemental material.

EFA parameters. Running EFA requires us to make informed choices regarding analysis parameters [108]. We selected the following analysis settings: (1) we used a Principal Axis (PA) factoring method, which does not entail distributional assumptions [34], because our data did not meet the normality requirements for Maximum Likelihood methods when we tested for univariate and multivariate normality; (2) an oblique rotation method (Promax) because underlying factors were likely to violate assumptions of independence, and (3) a common factor analysis model because it is better suited to scale development than Principal Component Analysis [30]. We also confirmed that our data's correlation matrix was factorable before running the EFA. We report results from all tests we conducted prior to EFA in Appx. I.

Number of factors. Following our reference literature recommendations [30, 108], we used *parallel analysis* and *scree plots* (Appx. J) to assess how many factors were likely to explain covariance patterns of our candidate items. In EFA, a parallel analysis determines the number of factors that capture more variance than what would be expected by random chance. In our case, 3–5 factors seemed to be necessary to explain the variance in our data, as we show in Table 2.

Parallel analysis, though, does not provide a definitive answer as to how many factors should be retained during scale development [30]. Instead, researchers should analyze the output *factor loadings* tables and assess whether or not groups of items appear to reflect meaningful constructs. Because scree plots showed a distinct slope break after factor 1, and parallel analysis suggested 3–5 factors depending on the data subsets, we examined factor loadings tables for structures ranging from 1 to 5 factors (more detail in Appx. K). We also examined *model fit metrics* for each factor structure [35], which we detail in our OSF Research log. From these combined observations, we concluded that a 4-factors solution produced the most meaningful grouping of items to describe perceived readability. We then removed items with cross-loadings (loadings in more than one factor with a value above a cutoff value of 0.32 [99]) and conducted a Multi-Group Confirmatory Factor Analysis (MG-CFA) to assess how appropriate the 4-factor solution was to explain variance in collected rating, across individual stimuli. We present a short description of this analysis in Appx. L. The resulting fit indices shown in Table 19 allowed us to confidently proceed with the development of PREVis as an instrument with four subscales.

Four factors in PREVis. As we show in Fig. 28 in Appx. L, we identified four underlying constructs from our EFA: ◆ **UNDERSTAND**, which relates to the ease of understanding visual encodings with 7 items such as "It is obvious for me how to read this visualization"; ◆ **LAYOUT**, with 4 items related to visual clarity such as "I don't find this visualization crowded", or "I don't find this visualization messy"; ◆ **DATAREAD**, in reference to Curcio's first level of "*reading the data*" [28], with 5 items such as "I can easily find specific elements in this visualization", or "I can easily retrieve information from this visualization"; and ◆ **DATAFEAT**, which relates to Curcio's "*reading between the data*" [28], with the 2 items in our questionnaire related to seeing "data features (for example, a minimum, or an outlier, or a trend)". For the remainder of our work, we considered each of these groups of items as individual scales, and we refer to them as *subscales*.

5.5.2 Item reduction for each subscale

In this step we evaluated the performance of individual items to identify the most appropriate ones to constitute each subscale. The goal of item reduction is to obtain a parsimonious final instrument, while still ensuring reliability of its measurements. The main indicator of reliability used in scale development is Cronbach's alpha, which estimates the scale's total variance attributable to a common source [30]. DeVellis and Thorpe [30] consider alpha values of 0.7–0.8 to be acceptable, and 0.8–0.9 to be very good. Above 0.9, researchers should consider shortening the list of items. With four subscales in our final instrument, our main goal was to obtain good reliability with a minimum length. We decided to first target a number of 3 items per scale, as a construct with fewer than 3 items is generally weak and unstable [26], and to add more items if needed until reaching an alpha coefficient >0.8. We had only 2 items to integrate in ◆ **DATAFEAT**, but this factor was nonetheless very stable for all EFAs we conducted with 3 or more factors.

We refer to Appx. N for more details on the item reduction. We obtained the final set of items summarized in Table 21 and conducted a Multi-Group Confirmatory Factor Analysis (MG-CFA) to assess how appropriate our scales are to capture information from the exploratory survey across stimuli. We obtained good reliability for each scale (0.88–0.92, Table 27) and fit metrics for the final model (Table 25).

6 PREVIS SUBSCALES VALIDATION

Scales need proper validation before they can be used as data collection instruments. Validation means verifying (a) that the scale's **dimensionality** (i. e., number of factors) remains the same with an independent sample of the population, (b) that the scales' **reliability** remains high for this independent sample, and (c) that the scales actually measure the intended **construct** (here, perceived readability via four subscales: ◆ **UNDERSTAND**, ◆ **LAYOUT**, ◆ **DATAREAD**, and ◆ **DATAFEAT**). To validate PREVis, we ran a fourth experiment, again pre-registered (`osf.io/yex32`) and IRB-approved (Inria COERLE, avis №2023-17).

6.1 Study design

Dimensionality and reliability tests for scale validation are run on data from an independent sample, using the final tool in a similar context. We thus designed a second crowd-sourced survey to administer our scale, similar to that of the exploratory study, with the following differences: we used 3 node-link graphs as stimuli (Fig. 5, top), all participants rated all stimuli, and we presented rating items for each subscale on a single screen. For details of our study design see Appx. P.

In addition, construct validity tests require supplemental measures and correlation tests on three criteria: ❶ **inter-subscales reliability:** our subscales' scores should highly and positively correlate among themselves, indicating the existence of a shared underlying construct in respondents (i. e., perceived readability); ❷ **discriminant validity:** our subscales' scores should not correlate with measures of a different, unrelated construct, measured with a similar method (i. e., in our, case, a 7-point Likert scale); and ❸ **convergent validity:** our subscales' scores should positively correlate with another readability-related indicator measured using a different method.

For discriminant validity, we chose to use measures of the "extraversion" personality trait, which we obtained using the corresponding 2 scale items from a validated 10-items version of the Big Five personality Inventory [84]. We chose this measure as it is performed on a

7-point Likert scale, and we have no reason to think that extraversion should correlate—positively or negatively—to perceived readability in visualization. For convergent validity, as we stated in Sect. 2, we do not have a validated instrument to collect objective or subjective measures of readability. Some instruments exist, however, that can produce metrics related to readability: in particular, algorithms can compute graph readability metrics [31] on node-link layouts. We generated 3 different node-link visualizations with a `D3.js` force-directed component to serve as stimuli in this survey. We then used Gove's `Greadability.js` library [43] (`github.com/rpgove/greadability`) for calculating graph layout metrics on each graph, as we document in Appx. P.

6.2 Survey administration

Participants. We targeted a sample size of ≥ 110 participants (10 per item in the final tool). We recruited participants from Prolific, who had to be fluent English speakers of legal age, excluding people who participated in our previous studies. Having noticed that our population in the exploratory survey was skewed towards men—even though women are more represented on Prolific—, we added a gender distribution criterion. Participants received a compensation of €11.52/h.

Procedure. Participants in this survey first answered a consent form and a question about color vision deficiency. Then, we gave them a brief explanation on how to read a node-link diagram in the context of the study. In the next part of the survey, we presented participants with 3 different node-link graphs, in random order. For each diagram, respondents answered 2 reading questions and rated the visualization using our 11 PREVis items with a labeled 7-point Likert scale. We presented PREVis items grouped by subscale; we randomized the order of subscales, and the order of items within them. Participants then had an option to leave an additional text comment. Finally, they answered the two extraversion items. We detail the procedure further in Appx. Q.

6.3 Survey results

We had provisioned extra Prolific budget in case answering the survey would take people longer than we expected, but participants actually completed the questionnaire a little faster than expected. This allowed us, as preregistered, to extend the study to more participants until budget exhaustion; as a result, we recruited a total of 148 participants. All participants passed our single comprehension check and at least 2 out of our 3 attention checks. As a result, we included ratings from all participants (ages: mean 28.4 years, SD 7 years; 48% female, 48% male, 3% non-binary, and 1% gender not disclosed; education: 3% secondary education, 26% high school diploma, 13% technical or community college, 38% undergraduate degree, 18% graduate degree, 3% doctorate); color vision deficiency: 4% yes).

Tests of dimensionality. To assess the dimensionality of PREVis we conducted a parallel analysis and a Multi-Group Confirmatory Factor Analysis on our collected data, following the same approach as described in Appx. L for scale development. We detail in Appx. R our findings that a 4-factors structure was appropriate, and that the subscales met criteria of good model fit from reference literature [50].

Tests of reliability. We computed alpha and omega coefficients for each subscale. For all scales, all coefficients were in the 0.87–0.96 range (see Table 31 in Appx. R), which qualifies as very good [30].

Tests of validity. To test the validity of our PREVis instrument, we followed recommendations from our reference literature to create a multi-trait multi-method (MTMM) matrix [12, 30].

A MTMM matrix represents the correlations in data collected from 3 different instruments for assessment of convergent and discriminant validity, as we described in Sect. 6.1. In Appx. R.2.3 we detail how we generated the composite MTMM matrix in Fig. 4, which confirmed all 3 criteria: inter-subscales reliability (positive and high correlations in ❶), discriminant validity (correlations close to 0 in ❷), and convergent validity (positive correlations in ❸). Finally, we plotted PREVis ◆◆◆◆ subscales' average ratings with 95% CI for each of the 3 stimuli (Fig. 5). For each ◆◆◆◆ subscale, ratings allowed to discriminate clearly between the three stimuli, in the expected order: A > B > C.

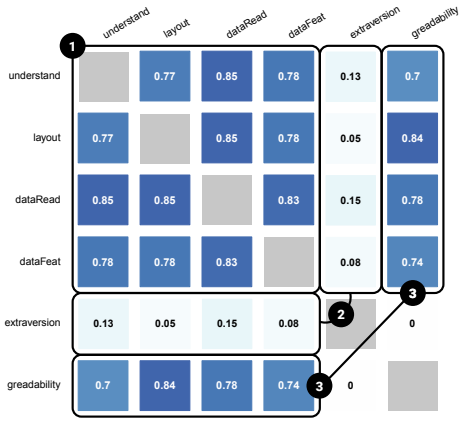

Fig. 4: Multi-trait multi-method composite correlation matrix (see details in Appx. R.2.3): ❶ reliability among PREVis subscales, ❷ discriminant validity from an unrelated personality trait in respondents, and ❸ convergent validity with graph layout metrics.

7 Discussion and Future Work

From our early investigations to define the readability domain, to our exploratory factor analysis in scale development, and our final evaluation of PREVis ◆◆◆◆ performance, we uncovered different components related to how people perceive readability of visualizations. This work allowed us to forge a strong opinion that readability cannot be expressed with a single aggregate measure. For example, when looking at ratings from our validation survey in Fig. 5, or from our exploratory survey in Fig. 29 in Appx. O, there are clear discrepancies between how easily people estimate they can ◆ **Understand** a visualization, and how clear they find that visualization's ◆ **Layout**. This is consistent with our view that three families of factors influence readability in data visualization (Fig. 2): the visualization *display*, the *reading task*, and the *reader*.

Items from ◆ **Understand**, in particular, appear to at least partly reflect a *reader's* familiarity with the displayed type of visualization, and by extension to their visualization literacy w.r.t. this specific type of representation [17]. It is particularly clear in ratings from our exploratory survey (see ratings in Fig. 29 in Appx. O and visualization stimuli in Fig. 10 in Appx. F): the unfamiliar GeneaQuilts visualization (stimulus E) received the lowest ◆ **Understand** score among all stimuli, while being rated higher in ◆ **Layout** than stimuli D (pie chart with gradient colors encodings) and F (line chart with 18 different categories). Future work should investigate the within-participant relationship between visualization literacy tests scores [27, 59] on certain visualizations idioms and ◆ **Understand** ratings on similar visualizations. ◆ **Layout** items, conversely, relate more to the clarity of the visualization *display* itself and, without surprise, showed the highest correlation with graph layout metrics among PREVis subscales during our validation study ($r = 0.74$ in Fig. 4). Future work should assess to what extent ◆ **Layout** alone can provide valid measures of display clarity in visualizations for which

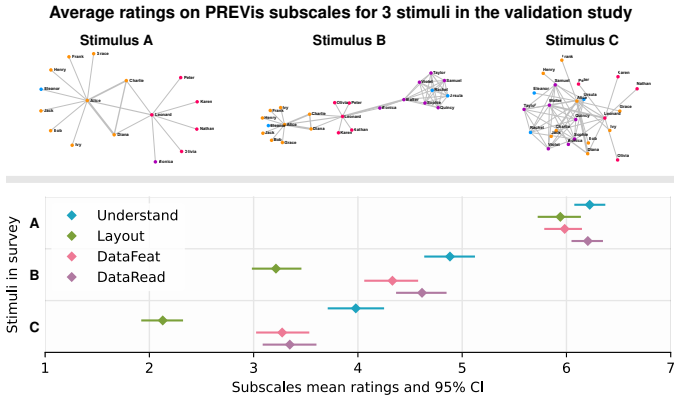

Fig. 5: Average ratings (from 1 = "Strongly disagree" to 7 = "Strongly agree") using the four PREVis ◆◆◆◆ subscales on three node-link visualizations of different readability levels (A > B > C).

we do not have layout metric algorithms. The two other dimensions in PREVis relate more to the outcomes of *reading tasks*: visual retrieval of data values in ◆ **DataRead**, and visual saliency of data patterns in ◆ **DataFeat**. These measures, when collected alongside task performance metrics such as speed and accuracy, can provide researchers with a more comprehensive view of how well a given visualization supported participants in their reading tasks. Future work should determine whether—and under which conditions—these indicators can predict actual reading task performance in respondents. If such work were to provide positive results, researchers might opt to use ◆ **DataRead** and ◆ **DataFeat** as proxies for reading task performance measures in studies where the visualization's objective efficiency is only a peripheral concern.

As such, each subscale in PREVis ◆◆◆◆ can be regarded as an individual measure, allowing researchers to choose a subset of the four scales in which they are particularly interested. We draw attention to the fact, however, that the collected data would then be missing the other dimensions we established, thus preventing researchers from drawing conclusions regarding *perceived readability*. It is our opinion that researchers who want to explore higher-level cognitive processes such as decision-making or acceptability of a new visualization system need to take perceived readability into account in its full dimensionality, among other variables. Future work could also examine the relationship between PREVis ◆◆◆◆ measurements with other aspects of visualization such as user engagement [87] or visualization trustworthiness [32]. Findings from such studies would allow us to expand our theoretical understanding of how people process information in visualizations, ultimately providing better guidance for visualization design and recommender systems.

8 Using PREVis ◆◆◆◆, Limitations, and Conclusion

When implementing PREVis in user studies, we recommend that participants at least attempt a few reading tasks on a visualization, before evaluating it with all items from each subscale ◆◆◆◆. They should answer items on a fully labeled 7-point Likert scale—at least, the neutral point and extremities must be labeled, and not numbered. We encourage researchers to provide an "I don't know / Not applicable" option or an optional comment field to let participants provide more detailed and qualitative information along with PREVis ◆◆◆◆ ratings. Indeed, in our two crowd-sourced surveys, we received many comments that were very relevant to better understanding our participants' reading experiences. We highlight that ◆ **Layout** items are negatively phrased, which we explain in Appx. E. We advise against reformulating these items—and PREVis items in general. However, researchers studying perceived visual clarity alone could choose to use only ◆ **Layout**, and reformulate those items' statements into affirmative sentences—in which case they should reverse scores before analyzing their data.

A limitation of this work is that we did not attempt to expose participants to more than one visualization before they started to rate images. While PREVis ◆◆◆◆ already demonstrated its ability to capture a broad spectrum of readability perceptions, presenting multiple designs prior to rating could potentially enhance the consistency of participants' ratings. On a related note, during our pre-test interviews we received suggestions from participants to use a readability comparison task (A *vs.* B) rather than a rating scale for a single visualization; similarly, in our validation study we received comments such as "This one was very messy, compared to the others", or "This representation is the best from far !!" We warn, however, against the temptation to adapt PREVis ◆◆◆◆ items as A/B testing questions because we do not know how reliable such measures would be. Perhaps even more importantly, as researchers have emphasized, generalization of visualization ranking is not without risks [29]. Converting PREVis into a ranking instrument would defeat another purpose of scales, which is to provide standardized results that facilitate cross-study comparisons and meta-analysis.

We developed PREVis as a versatile instrument for diverse visualization types and readers. As the low amount of unanswered questions in our exploratory survey demonstrates (Sect. 5.5.1), respondents were able to use PREVis in multiple visualization situations: familiar and unfamilar visualization idioms, categorical, ordinal and continuous numerical data types, a wide range of numbers of data entities, 1–4 encoded data attributes ... We acknowledge, however, that some visualization contexts may require more specific questions, and that researchers could find that our items are too vaguely phrased to accurately capture perceived readability in their specific research setting. In a network topology reading situation, for instance, one could find it useful to replace ◆"I can easily find specific elements in this visualization" with "I can easily find specific nodes in this graph"; or, when working with expert users on cluster detection in graphs, to replace the ◆"data features" expression with "data clusters." When doing so, researchers should follow methodological guidelines in scale development and validation—as we extensively documented in this work—to ensure that they produce a reliable and accurate measuring instrument.

Finally, we only focused on static 2D representations without any interaction. Readability is already complex to try to define in such a constrained set of situations; but once it is better established, future research work should expand our understanding of readability to 3D environments, in motion situations [112], and with interaction.

In conclusion, our PREVis ◆◆◆◆ instrument is readily deployable in many visualization contexts and, based on our validation, can provide reliable and nuanced measures of perceived readability to researchers. Our work showed that the construct of readability, as visualization readers perceive it, is too complex to be expressed in a single or aggregate measure, and we identified at least four subcomponents. As such, PREVis ◆◆◆◆ also provides multiple opportunities for future research, which can ultimately contribute to better explaining factors and outcomes of readability in data visualization.

Acknowledgments

We thank Markus Wallinger, who provided their material from [105], as well as all Aviz team members for their feedback on the project.

Supplemental Material Pointers

All supplemental materials are available on OSF at `osf.io/9cg8j`. We also share our code at `github.com/AF-Cabouat/PREVis-scales`.

Figure Credits and Copyright

Fig. 2 was adapted from [37,48] but is a full re-creation. All our figures remain under our own copyright and are available under the Creative Commons CC BY 4.0 license; we share them at `osf.io/9cg8j`.

References

[1] K. Ajani, E. Lee, C. Xiong, C. N. Knaflic, W. Kemper, and S. Franconeri. Declutter and Focus: Empirically Evaluating Design Guidelines for Effective Data Communication. *IEEE Trans Vis Comput Graph*, 28(10):3351–3364, 2022. doi: 10/gn984r

[2] R. Alroobaea and P. J. Mayhew. How many participants are really enough for usability studies? In *Proc. SAI*, pp. 48–56. SAI, Cleckheaton, 2014. doi: 10/gtgz8q

[3] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *Proc. InfoVis*, pp. 111–117. IEEE CS, Los Alamitos, 2005. doi: 10/bwrm27

[4] C. Bachmaier. A radial adaptation of the Sugiyama framework for visualizing hierarchical information. *IEEE Trans Vis Comput Graph*, 13(3):583–594, 2007. doi: 10/dr27xr

[5] A. Bailin and A. Grafstein. The linguistic assumptions underlying readability formulae: A critique. *Lang Commun*, 21(3):285–301, 2001. doi: 10/d6jngm

[6] K. Börner, A. Maltese, R. N. Balliet, and J. Heimlich.Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors. *Inf Vis*, 15(3):198–213, 2016. doi: 10/f8sgws

[7] P. C. Beatty and G. B. Willis.Research synthesis: The practice of cognitive interviewing. *Public Opin Q*, 71(2):287–311, 2007. doi: 10/b7pdj4

[8] R. G. Benjamin. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educ Psychol Rev*, 24(1):63–88, 2012. doi: 10/bdjfkd

[9] C. Bennett, J. Ryall, L. Spalteholz, and A. Gooch.The aesthetics of graph visualization. In *Proc. CAe*, pp. 57–64. EG, Goslar, 2007. doi: 10/gn9kwr

[10] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. Esri Press, Redlands, California, 1983.

[11] A. Bezerianos, P. Dragicevic, J.-D. Fekete, J. Bae, and B. Watson. GeneaQuilts: A system for exploring large genealogies. *IEEE Trans Vis Comput Graph*, 16(6):1073–1081, 2010. doi: 10/btb4js

[12] G. O. Boateng, T. B. Neilands, E. A. Frongillo, H. R. Melgar-Quiñonez, and S. L. Young. Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Front Public Health*, 6, art. no. 149, 18 pages, 2018. doi: 10/gfsqzs

[13] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete. A principled way of assessing visualization literacy. *IEEE Trans Vis Comput Graph*, 20(12):1963–1972, 2014. doi: 10/f6qjv6

[14] R. Brath. Metrics for effective information visualization. In *Proc. InfoVis*, pp. 108–111. IEEE CS, Los Alamitos, 1997. doi: 10/dm997r

[15] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Trans Vis Comput Graph*, 19(12):2376–2385, 2013. doi: 10/f5h3q4

[16] C. Bu, Q. Zhang, Q. Wang, J. Zhang, M. Sedlmair, O. Deussen, and Y. Wang. SineStream: Improving the readability of streamgraphs by minimizing sine illusion effects. *IEEE Trans Vis Comput Graph*, 27(2):1634–1643, 2021. doi: 10/ghv58n

[17] A.-F. Cabouat, T. He, F. Cabric, T. Isenberg, and P. Isenberg. Position paper: A case to study the relationship between data visualization readability and visualization literacy. In *Proc. CHI Workshop "Toward a More Comprehensive Understanding of Visualization Literacy"*, 2024. Online: hal.science/hal-04523790.

[18] A.-F. Cabouat, T. He, P. Isenberg, and T. Isenberg. Pondering the reading of visual representations, 2023. Online: hal.science/hal-04240900.

[19] Readability, n. In *Cambridge Advanced Learner's Dictionary & Thesaurus*. Cambridge University Press, 2021.

[20] S. K. Card, J. D. Mackinlay, and B. Shneiderman, eds. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco, 1999.

[21] Y.-H. Chan, C. D. Correa, and K.-L. Ma. The generalized sensitivity scatterplot. *IEEE Trans Vis Comput Graph*, 19(10):1768–1781, 2013. doi: 10/f47sfd

[22] S. Y. Y. Chyung, K. Roberts, I. Swanson, and A. Hankinson. Evidence-based survey design: The use of a midpoint on the Likert scale. *Perform Improv*, 56(10):15–23, 2017. doi: 10/gfgm4w

[23] L. A. Clark and D. Watson. Constructing validity: Basic issues in objective scale development. *Psychol Assess*, 7(3):309–319, 1995. doi: 10/bmw7mm

[24] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J Am Stat Assoc*, 79(387):531–554, 1984. doi: 10/gdvmwd

[25] F. Conrad and J. Blair. From impressions to data increasing the objectivity of cognitive interviews. In *JSM Proceedings*, pp. 1–9. ASA, Alexandria, 1996. Online: asasrms.org/Proceedings/papers/1996_001.pdf.

[26] A. B. Costello and J. Osborne. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pract Assess Res Eval*, 10, art. no. 7, 9 pages, 2005. doi: 10/ghgv6m

[27] Y. Cui, L. W. Ge, Y. Ding, F. Yang, L. Harrison, and M. Kay. Adaptive assessment of visualization literacy. *IEEE Trans Vis Comput Graph*, 30(1):628–637, 2024. doi: 10/gtjwqh

[28] F. R. Curcio. Comprehension of mathematical relationships expressed in graphs. *J Res Math Educ*, 18(5):382–393, 1987. doi: 10/fjh8jc

[29] R. Davis, X. Pu, Y. Ding, B. D. Hall, K. Bonilla, M. Feng, M. Kay, and L. Harrison. The risks of ranking: Revisiting graphical perception to model individual differences in visualization performance. *IEEE Trans Vis Comput Graph*, 30(3):1756–1771, 2024. doi: 10/gtjwqp

[30] R. F. DeVellis and C. T. Thorpe. *Scale Development: Theory and Applications*. SAGE, 5th ed., 2021. urn: urn:oclc:record:1245766436.

[31] C. Dunne, S. I. Ross, B. Shneiderman, and M. Martino. Readability metric feedback for aiding node-link visualization designers. *IBM J Res Dev*, 59(2/3), art. no. 14, 16 pages, 2015. doi: 10/gtn9pf

[32] H. Elhamdadi, A. Stefkovics, J. Beyer, E. Moerth, H. Pfister, C. X. Bearfield, and C. Nobre. Vistrust: A multidimensional framework and empirical study of trust in data visualizations. *IEEE Trans Vis Comput Graph*, 30(1):348–358, 2024. doi: 10/gtjwqg

[33] A. M. Elleman and E. L. Oslund. Reading comprehension research: Implications for practice and policy. *Policy Insights Behav Brain Sci*, 6(1):3–11, 2019. doi: 10/gmcgwb

[34] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychol Methods*, 4(3):272–299, 1999. doi: 10/b2ztct

[35] W. H. Finch. Using fit statistic differences to determine the optimal number of factors to retain in an exploratory factor analysis. *Educ Psychol Meas*, 80(2):217–241, 2020. doi: 10/ggjw7k

[36] E. E. Firat, A. Joshi, and R. S. Laramee. Interactive visualization literacy: The state-of-the-art. *Inf Vis*, 21(3):285–310, 2022. doi: 10/gpngsr

[37] A. R. Fox. Theories and models in graph comprehension. In *Visualization Psychology*, pp. 39–64. Springer, Cham, 2023. doi: 10/gtgz78

[38] E. G. Freedman and P. Shah. Toward a model of knowledge-based graph comprehension. In *Proc. Diagrams*, pp. 18–30. Springer, Berlin, 2002. doi: 10/fr56t9

[39] S. N. Friel, F. R. Curcio, and G. W. Bright. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *J Res Math Educ*, 32(2):124–158, 2001. doi: 10/bmpjc8

[40] M. Ghoniem, J.-D. Fekete, and P. Castagliola. On the readability of graphs using node-link and matrix-based representations: A controlled experiment and statistical analysis. *Inf Vis*, 4(2):114–135, 2005. doi: 10/bncqv6

[41] L. Giovannangeli, F. Lalanne, R. Giot, and R. Bourqui. Guaranteed visibility in scatterplots with tolerance. *IEEE Trans Vis Comput Graph*, 30(1):792–802, 2024. doi: 10/gtgz8h

[42] P. Goffin, W. Willett, J.-D. Fekete, and P. Isenberg. Exploring the placement and design of word-scale visualizations. *IEEE Trans Vis Comput Graph*, 20(12):2291–2300, 2014. doi: 10/f6qjwg

[43] R. Gove. It pays to be lazy: Reusing force approximations to compute better graph layouts faster. OSF preprint, 2018. doi: 10/gftd8t

[44] H. Haleem, Y. Wang, A. Puri, S. Wadhwa, and H. Qu. Evaluating the readability of force directed graph layouts: A deep learning approach. *IEEE Comput Graph Appl*, 39(4):40–53, 2019. doi: 10/gndwtv

[45] K. W. Hall, A. Kouroupis, A. Bezerianos, D. A. Szafir, and C. Collins. Professional differences: A comparative study of visualization task performance and spatial ability across disciplines. *IEEE Trans Vis Comput Graph*, 28(1):654–664, 2022. doi: 10/gtgz8b

[46] T. He, P. Isenberg, R. Dachselt, and T. Isenberg. BeauVis: A validated scale for measuring the aesthetic pleasure of visual representations. *IEEE Trans Vis Comput Graph*, 29(1):363–373, 2023. doi: 10/kt3n

[47] C. Healey and J. Enns. Attention and visual memory in visualization and computer graphics. *IEEE Trans Vis Comput Graph*, 18(7):1170–1188, 2012. doi: 10/ch7sc2

[48] M. Hegarty. The cognitive science of visual-spatial displays: Implications for design. *Top Cognit Sci*, 3(3):446–474, 2011. doi: 10/c3274w

[49] N. Henry, A. Bezerianos, and J.-D. Fekete. Improving the readability of clustered social networks using node duplication. *IEEE Trans Vis Comput Graph*, 14(6):1317–1324, 2008. doi: 10/bd4x4k

[50] L.-t. Hu and P. M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Struct Equation Model*, 6(1):1–55, 1999. doi: 10/dbt

[51] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller. A systematic review on the practice of evaluating visualization. *IEEE Trans Vis Comput Graph*, 19(12):2818–2827, 2013. doi: 10/f5h29z

[52] M. Koch, K. Kurzhals, M. Burch, and D. Weiskopf. Visualization psychology for eye tracking evaluation. In *Visualization Psychology*, pp. 243–260. Springer, Cham, 2023. doi: 10/gtgz8m

[53] R. Kosara. Visualization criticism – The missing link between information visualization and art. In *Proc. IV*, pp. 631–636. IEEE CS, Los Alamitos, 2007. doi: 10/dm75nn

[54] R. Kosara. An empire built on sand: Reexamining what we think we know about visualization. In *Proc. BELIV*, pp. 162–168. ACM, New York, 2016. doi: 10/gfz5kr

[55] S. M. Kosslyn. Understanding charts and graphs. *Appl Cognit Psychol*, 3(3):185–225, 1989. doi: 10/cvdzr5

[56] K. Kurzhals, B. Fisher, M. Burch, and D. Weiskopf. Evaluating visual analytics with eye tracking. In *Proc. BELIV*, pp. 61–69. ACM, New York, 2014. doi: 10/gtgz75

[57] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Trans Vis Comput Graph*, 18(9):1520–1536, 2012. doi: 10/drrh6j

[58] S. Lee, S.-H. Kim, Y.-H. Hung, H. Lam, Y.-A. Kang, and J. S. Yi. How do people make sense of unfamiliar visualizations?: A grounded model of novice's information visualization sensemaking. *IEEE Trans Vis Comput Graph*, 22(1):499–508, 2016. doi: 10/gfw4vs

[59] S. Lee, S.-H. Kim, and B. C. Kwon. VLAT: Development of a visualization literacy assessment test. *IEEE Trans Vis Comput Graph*, 23(1):551–560, 2017. doi: 10/f92d38

[60] Z. Li, R. Shi, Y. Liu, S. Long, Z. Guo, S. Jia, and J. Zhang. Dual space coupling model guided overlap-free scatterplot. *IEEE Trans Vis Comput Graph*, 29(1):657–667, 2023. doi: 10/gtgz8j

[61] R. J. Little. A test of missing completely at random for multivariate data with missing values. *J Am Stat Assoc*, 83(404):1198–1202, 1988. doi: 10/c43d3q

[62] Z. Liu, R. J. Crouser, and A. Ottley. Survey on individual differences in visualization. *Comput Graph Forum*, 39(3):693–712, 2020. doi: 10/gg6cr5

[63] M. A. Livingston, L. E. Matzen, A. Harrison, A. Lulushi, M. Daniel, M. Dass, D. Brock, and J. W. Decker. A study of perceptual and cognitive models applied to prediction of eye gaze within statistical graphs. In *Proc. SAP*, art. no. 6, 9 pages. ACM, New York, 2020. doi: 10/gpjc7n

[64] J. Mackinlay. Automating the design of graphical presentations of relational information. *ACM Trans Graph*, 5(2):110–141, 1986. doi: 10/dxdkkp

[65] L. E. Matzen, M. J. Haass, K. M. Divis, and M. C. Stites. Patterns of attention: How data visualizations are read. In *Proc. AC*, pp. 176–191. Springer, Cham, 2017. doi: 10/gtgz8n

[66] L. E. Matzen, M. J. Haass, J. Tran, L. A. McNamara, M. J. Haass, J. Tran, and L. A. McNamara. Using eye tracking metrics and visual saliency maps to assess image utility. *Electron Imaging*, 28, art. no. art00033, 8 pages, 2016. doi: 10/gtgz8k

[67] D. McNeish. Exploratory factor analysis with small samples and missing data. *J Personality Assess*, 99(6):637–652, 2017. doi: 10/gfvqgr

[68] B. J. F. Meyer. Text coherence and readability. *Top Lang Disord*, 23(3):204, 2003. doi: 10/bgcx6k

[69] A. Mirzaei, S. R. Carter, A. E. Patanwala, and C. R. Schneider. Missing data in surveys: Key concepts, approaches, and applications. *Res Social Administrative Pharm*, 18(2):2308–2316, 2022. doi: 10/ktrm

[70] T. Munzner. *Visualization Analysis and Design*. CRC Press, Boca Raton, 2014. doi: 10/gd3xgq

[71] V. Nassiri, A. Lovik, G. Molenberghs, and G. Verbeke. On using multiple imputation for exploratory factor analysis of incomplete data. *Behav Res Methods*, 50(2):501–517, 2018. doi: 10/gf6rz3

[72] R. Netzel, B. Ohlhausen, K. Kurzhals, R. Woods, M. Burch, and D. Weiskopf. User performance and reading strategies for metro maps: An eye tracking study. *Spatial Cognit Comput*, 17(1):39–64, 2017. doi: 10/gtgz8p

[73] S. Nusrat, M. J. Alam, and S. Kobourov. Evaluating Cartogram Effectiveness. *IEEE Trans Vis Comput Graph*, 24(2):1077–1090, 2018. doi: 10/gct24q

[74] L. M. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. Decision making with visualizations: A cognitive framework across disciplines. *Cognit Res Princ Implic*, 3(1), art. no. 29, 25 pages, 2018. doi: 10/ggrtng

[75] E. M. Peck, B. F. Yuksel, L. Harrison, A. Ottley, and R. Chang. Towards a 3-dimensional model of individual cognitive differences: Position paper. In *Proc. BELIV*, art. no. 6, 6 pages. ACM, 2012. doi: 10/ggjbvc

[76] C. Perfetti and J. Stafura. Word knowledge in a theory of reading comprehension. *Scie Stud Reading*, 18(1):22–37, 2014. doi: 10/gf3bmc

[77] C. Perin, J. Boy, and F. Vernier. Using gap charts to visualize the temporal evolution of ranks and scores. *IEEE Comput Graph Appl*, 36(5):38–49, 2016. doi: 10/gthk98

[78] S. Pinker. A theory of graph comprehension. In *Artificial Intelligence and the Future of Testing*, chap. 4, pp. 73–126. Lawrence Erlbaum Assoc., Hillsdale, 1990. urn: urn:oclc:record:1148020681.

[79] P. Polatsek, M. Waldner, I. Viola, P. Kapec, and W. Benesova. Exploring visual attention and saliency modeling for task-based visual analysis. *Comput Graph*, 72:26–38, 2018. doi: 10/gdhtkf

[80] M. Porter. An algorithm for suffix stripping. *Program Electron Lib Inf Syst*, 14(3):130–137, 1980. doi: 10/dnzbpm

[81] Y. Postigo and J. I. Pozo. On the road to graphicacy: The learning of graphical representation systems. *Educ Psychol*, 24(5):623–644, 2004. doi: 10/c5395g

[82] H. C. Purchase. Metrics for graph drawing aesthetics. *J Visual Lang Comput*, 13(5):501–516, 2002. doi: 10/fdqd3w

[83] G. J. Quadri and P. Rosen. A survey of perception-based visualization studies by task. *IEEE Trans Vis Comput Graph*, 28(12):5026–5048, 2022. doi: 10/gr6323

[84] B. Rammstedt and O. P. John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in English and German. *J Res Personality*, 41(1):203–212, 2007. doi: 10/djzd32

[85] P. Ruchikachorn and K. Mueller. Learning visualizations by analogy: Promoting visual literacy through visualization morphing. *IEEE Trans Vis Comput Graph*, 21(9):1028–1044, 2015. doi: 10/ggjbtd

[86] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE Trans Vis Comput Graph*, 20(12):1604–1613, 2014. doi: 10/f6qj6x

[87] B. Saket, A. Endert, and J. Stasko. Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proc. BELIV*, pp. 133–142. ACM, New York, 2016. doi: 10/gfw4vt

[88] M. Sedlmair, A. Frank, T. Munzner, and A. Butz. RelEx: Visualization for actively changing overlay network specifications. *IEEE Trans Vis Comput Graph*, 18(12):2729–2738, 2012. doi: 10/f4ft3t

[89] P. Shah, E. G. Freedman, and I. Vekiri. The comprehension of quantitative information in graphical displays. In *The Cambridge Handbook of Visuospatial Thinking*, chap. 11, pp. 426–476. Cambridge University Press, 2005. doi: 10/crwhhf

[90] J. A. Shepperd, G. Pogge, J. M. Hunleth, S. Ruiz, and E. A. Waters. Guidelines for conducting virtual cognitive interviews during a pandemic. *J Med Internet Res*, 23(3), art. no. e25173, 5 pages, 2021. doi: 10/gq6vnc

[91] S. Shin, S. Chung, S. Hong, and N. Elmqvist. A Scanner Deeply: Predicting gaze heatmaps on visualizations using crowdsourced eye movement data. *IEEE Trans Vis Comput Graph*, 29(1):396–406, 2023. doi: 10/mndf

[92] D. Simkin and R. Hastie. An information-processing analysis of graph perception. *J Am Stat Assoc*, 82(398):454–465, 1987. doi: 10/gdvmzq

[93] D. Skau and R. Kosara. Readability and precision in pictorial bar charts. In *EuroVis Short Papers*, pp. 91–95. EG, Goslar, 2017. doi: 10/gtgz8f

[94] S. Smart and D. A. Szafir. Measuring the separability of shape, size, and color in scatterplots. In *Proc. CHI*, art. no. 669, 14 pages. ACM, New York, 2019. doi: 10/gf2b87

[95] M. Solen. Scoping the future of visualization literacy: A review. In *Proc. VisComm*, art. no. 2, 6 pages, 2022. doi: 10/gtgz79

[96] L. South, D. Saffo, O. Vitek, C. Dunne, and M. A. Borkin. Effective use of Likert scales in visualization evaluations: A systematic review. *Comput Graph Forum*, 41(3):43–55, 2022. doi: 10/m5t6

[97] J. Stasko. Value-driven evaluation of visualizations. In *Proc. BELIV*, pp. 46–53. ACM, New York, 2014. doi: 10/gtgz74

[98] B. Steichen, G. Carenini, and C. Conati.User-adaptive information visualization: Using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proc. IUI*, pp. 317–328. ACM, New York, 2013. doi: 10/gtgz8c

[99] H. Taherdoost, S. Sahibuddin, and N. Jalaliyoon. Exploratory factor analysis; Concepts and theory. In *Proc. MCSS*, pp. 375–382. WSEAS, 2014. Online: ssrn.com/abstract=4178683.

[100] The reVISit team. reVISit: Scalable empirical evaluation of interactive visualizations. Web site: revisit.dev, 2022. Visited March 2024.

[101] A. Thudt, J. Walny, C. Perin, F. Rajabiyazdi, L. MacDonald, R. Vardeleon, S. Greenberg, and S. Carpendale. Assessing the readability of stacked graphs. In *Proc. GI*, pp. 167–174. CHCCS, Waterloo, 2016. doi: 10/gtgz8d

[102] S. B. Trickett and J. G. Trafton. Toward a comprehensive model of graph comprehension: Making the case for spatial cognition. In *Proc. Diagrams*, pp. 286–300. Springer, Berlin, 2006. doi: 10/dc4f78

[103] Y. Tu and H.-W. Shen. Visualizing changes of hierarchical data using treemaps. *IEEE Trans Vis Comput Graph*, 13(6):1286–1293, 2007. doi: 10/fq2wwj

[104] D. S. Valdivia and S. Dai. Number of response categories and sample size requirements in polytomous IRT models. *J Exp Educ*, 92(1):154–185, 2024. doi: 10/gtn9pd

[105] M. Wallinger, B. Jacobsen, S. Kobourov, and M. Nöllenburg. On the readability of abstract set visualizations. *IEEE Trans Vis Comput Graph*, 27(6):2821–2832, 2021. doi: 10/gtgz8g

[106] J. Wang, X. Cai, J. Su, Y. Liao, and Y. Wu. What makes a scatterplot hard to comprehend: Data size and pattern salience matter. *J Vis*, 25(1):59–75, 2022. doi: 10/gtjwph

[107] C. Ware, H. Purchase, L. Colpoys, and M. McGill. Cognitive measurements of graph aesthetics. *Inf Vis*, 1(2):103–110, 2002. doi: 10/dp72pq

[108] M. W. Watkins. Exploratory factor analysis: A guide to best practice. *J Black Psychol*, 44(3):219–246, 2018. doi: 10/gdk2zx

[109] Z. Wen, Y. Liu, S. Tan, J. Chen, M. Zhu, D. Han, J. Yin, M. Xu, and W. Chen. Quantivine: A visualization approach for large-scale quantum circuit representation and analysis. *IEEE Trans Vis Comput Graph*, 30(1):573–583, 2024. doi: 10/gt2v32

[110] C. Xiong, C. Stokes, Y.-S. Kim, and S. Franconeri. Seeing what you believe or believing what you see? Belief biases correlation estimation. *IEEE Trans Vis Comput Graph*, 29(1):493–503, 2023. doi: 10/gtg7mh

[111] M. A. Yalçin, N. Elmqvist, and B. B. Bederson. Cognitive stages in visual data exploration. In *Proc. BELIV*, pp. 86–95. ACM, New York, 2016. doi: 10/gfz686

[112] L. Yao, A. Bezerianos, R. Vuillemot, and P. Isenberg. Visualization in motion: A research agenda and two evaluations. *IEEE Trans Vis Comput Graph*, 28(10):3546–3562, 2022. doi: 10/gr633b

[113] Y. Zhang, X. Ding, and N. Gu. Understanding fatigue and its impact in crowdsourcing. In *Proc. CSCWD*, pp. 57–62. IEEE CS, Los Alamitos, 2018. doi: 10/gm248k

PREVis: Perceived Readability Evaluation for Visualizations

Appendix

In this appendix we provide additional explanations, tables, plots, and charts that show data beyond the material that we could include in the main paper due to space limitations or because it was not essential for explaining our approach.

CONTENTS

A SPECIFIC GLOSSARY

In the item development section as well as elsewhere in the paper, we use the following terms for which we provide here our contextualized definition to clarify the discussion better:

- **Factor:** we use this word in the context of Exploratory and Confirmatory Factor Analyses where it refers to a latent variable that explains a cluster of covariance among the observed variables (in our case, items). Mathematically, factors are associated with the eigenvectors of a correlation matrix, which we derived from item responses. Each factor is also linked to an eigenvalue, which indicates the proportion of variance in the observed data that the factor accounts for. Within a factor, items have coefficients (i. e., "loadings") that represent the degree to which each item is associated with that factor. High loadings on a factor suggest that the items are strongly related to the factor's latent variable. By examining the content and nature of the items, we can interpret the meaning of the factor and propose a name for the underlying latent variable it represents. This how we determined the dimensions in PREVis ◆◆◆◆ as we detail in Appx. K.
- **Fully-labeled Likert scale:** in the context of items answer options, a fully-labeled Likert scale is a rating scale where each point is associated with a descriptor, also called a *text anchor* [96].
- **Instrument:** a measuring tool. In our case, the PREVis instrument consists of a group of 4 related scales.
- **Instrument dimensions:** dimensions are distinct components that contribute to the full construct being assessed (i. e., perceived readability). In our case, each dimension is measured using a dedicated *subscale*: ◆ **UNDERSTAND**: the intelligibility of the encodings for the reader; ◆ **LAYOUT**: the visual clarity of the layout; ◆ **DATAREAD**: how easily people feel they can read data values; and ◆ **DATAFEAT**: how easily people feel they can read data patterns. We derived these dimensions from the four *factors* (see above) we found during our Exploratory Factor Analysis described in Sect. 5.5.
- **Instrument validity:** validated instruments have been tested to verify that they measure the target construct (i. e., in our case, perceived readability). Validity is established throughout the process of developping an instrument [12].
- **Item:** in the context of a scale, an item is the combination of a statement or a question with its accompanying answer options. In our work, all items share the same 7-point rating scale answer options.
- **Perceived readability:** how readable a person finds a specific visualization in a given context. It is the construct that PREVis targets in respondents as a measuring *instrument*.
- **Reliability:** a scale's reliability is an indicator of its consistency and stability in measuring what it is intended to measure. Common indices of reliability, such as Cronbach's alpha and McDonald's omega, are based on correlation estimates among the

items within the scale. These indices assess the internal consistency, reflecting how well the items correlate with each other and consistently measure the same underlying construct.

- **Psychological scale:** an instrument measuring a single construct in respondents (i. e., how readable people find a visualization). Scales are generally used to help evaluate latent variables, i. e., traits or constructs for which direct observation is not possible. Therefore, scale are considered to be *indicators* of the target variable.
- **Subscale:** in this paper, we call "subscales" the four scales that form PREVis ◆◆◆◆, each measuring a specific dimension of perceived readability. We use the term subscale for the sake of readability in the paper. It is, however, different from the usual acceptance of the term in psychology research, where subscales' scores are usually aggregated to form a higher-level score. For PREVis ◆◆◆◆, we advise against this practice (see the discussion in Sect. 7 as well as in our practical companion PDF on `osf.io/9cg8j`).
- **Term:** a keyword we collected in one of our source corpora.

B Additional Term Collection Details

B.1 Deductive method sources

As described in Sect. 4.1.1, we extracted terms from study questionnaires (Pool 1) and participant comments (Pool 2). Table 3 shows all 55 sources in this work, and whether they contributed to Pool 1, or Pool 2, or both. 13 sources were common to the two pools of terms.

Table 4 shows the list of studies and collected questions from which we extracted terms to form Pool 1 in Sect. 4.1.1. We extracted all words from the questions as terms, except for stop words from the `nltk` package in `Python` (e. g., "to", "of", "for", "between").

Table 5 shows the list of studies with user comments from which we manually extracted terms to form Pool 2 in Sect. 4.1.1, along with the list of collected terms.

B.2 Inductive method sources

We provide the table of term collection from Sect. 4.1.2 as a separate file in our supplemental material because it contains 152 lines, which makes it too long for comfortable reading from a paginated PDF. To form this pool of terms, we extracted individual words from a selection of key expressions. Therefore, the column "key_expressions" hold the collected terms in the reference file.

We also provide a printout of the survey we used to collect statements from experts as supplemental material.

B.3 Combining all terms pools

Table 6 shows all collected terms from the 3 different pools of items described in Sect. 4.1.

Error log. In both Table 4 and Table 6 we noticed a term which should have been excluded: "visualization". Its presence is the result of a manual copy error, which we noticed too late in our qualitative analysis process for correcting the data. As all other instances of this word were initially excluded, the associated count do not reflect the actual content of the analyzed questionnaires.

Table 3: 55 publications from which we extracted terms for Pool 1 and Pool 2.

DOI	Pool 1 (studies)	Pool2 (comments)
10.1109/TVCG.2013.151	Yes	Yes
10.1109/TVCG.2012.189	Yes	Yes
10.1109/TVCG.2015.2467872	Yes	Yes
10.1109/TVCG.2021.3068337	Yes	Yes
10.1109/TVCG.2017.2745941	Yes	Yes
10.1109/TVCG.2022.3209475	Yes	Yes
10.1109/TVCG.2020.3030358	Yes	Yes
10.1109/TVCG.2012.255	Yes	Yes
10.1109/TVCG.2018.2865192	Yes	Yes
10.1109/TVCG.2020.3030388	Yes	Yes
10.1109/TVCG.2011.186	Yes	Yes
10.1109/TVCG.2017.2744118	Yes	Yes
10.1109/TVCG.2018.2835485	Yes	Yes
10.1109/TVCG.2021.3114789	Yes	-
10.1109/TVCG.2011.183	Yes	-
10.1109/TVCG.2022.3144975	Yes	-
10.1109/TVCG.2022.3163727	Yes	-
10.1109/TVCG.2020.3030404	Yes	-
10.1109/TVCG.2015.2467035	Yes	-
10.1109/TVCG.2021.3092680	Yes	-
10.1109/TVCG.2010.194	Yes	-
10.1109/MCG.2016.100	Yes	-
10.1109/TVCG.2014.2346983	Yes	-
10.1109/TVCG.2020.3004137	Yes	-
10.1109/TVCG.2012.225	Yes	-
10.1109/TVCG.2022.3209354	Yes	-
10.1109/TVCG.2016.2642109	Yes	-
10.1109/TVCG.2020.3030437	Yes	-
10.1109/TVCG.2013.180	Yes	-
10.1109/TVCG.2018.2865232	Yes	-
10.1109/TVCG.2018.2865049	Yes	-
10.1109/TVCG.2021.3085327	Yes	-
10.1109/TVCG.2019.2941208	Yes	-
10.1109/TVCG.2011.193	Yes	-
10.1109/TVCG.2021.3114775	-	Yes
10.1109/TVCG.2013.76	-	Yes
10.1109/TVCG.2022.3209480	-	Yes
10.1109/TVCG.2021.3114822	-	Yes
10.1109/TVCG.2022.3209484	-	Yes
10.1109/TVCG.2014.2346420	-	Yes
10.1109/TVCG.2013.191	-	Yes
10.1109/TVCG.2022.3209477	-	Yes
10.1109/TVCG.2014.2329308	-	Yes
10.1109/TVCG.2019.2934784	-	Yes
10.1109/TVCG.2009.176	-	Yes
10.1109/INFVIS.2002.1173148	-	Yes
10.1109/TVCG.2019.2934557	-	Yes
10.1109/TVCG.2020.2968911	-	Yes
10.1109/TVCG.2014.2337337	-	Yes
10.1109/TVCG.2019.2934337	-	Yes
10.1109/TVCG.2019.2934669	-	Yes
10.1109/TVCG.2013.233	-	Yes
10.1109/TVCG.2019.2906900	-	Yes
10.1109/VAST.2009.5332595	-	Yes
10.1109/TVCG.2018.2864907	-	Yes
10.1109/TVCG.2011.160	-	Yes

Table 4: 34 studies with questionnaires from which we extracted terms in Pool 1 (see Sect. 4.1.1).

DOI	Year	Type of visualization	Collected questions from which we extracted terms for item development
10.1109/TVCG.2022.3175626	2010	trends	readability; ease of use; visually cluttered
10.1109/TVCG.2022.3175626	2011	node-links; maps	confidence; effectiveness; frustration; ease; clutter; enjoyment; simplicity; ease to follow
10.1109/TVCG.2022.3175626	2011	node-links	intuitive
10.1109/TVCG.2022.3175626	2011	paths	clearly visible; visual clutter
10.1109/TVCG.2022.3175626	2012	matrix; node-links	clear arrangement
10.1109/TVCG.2022.3175626	2012	flows; paths	easy / hard to interpret; easy / hard to understand
10.1109/TVCG.2022.3175626	2012	node-links	effectiveness; prefered look
10.1109/TVCG.2022.3175626	2013	node-links	easy to learn to read; confident; low clutter
10.1109/TVCG.2022.3175626	2013	location; color	easy to gain a good overview of data; easy to understand displayed information; easy to interpret data values (in visualization); confidence in correct answers
10.1109/TVCG.2022.3175626	2014	nan	easy / difficult
10.1109/TVCG.2022.3175626	2015	diagrams	easy to read text; easy to read symbols; easy to see link between components; good overview of diagram in mind
10.1109/TVCG.2022.3175626	2015	parallel plots	easy to understand
10.1109/TVCG.2022.3175626	2016	line charts	easy / difficult to identify
10.1109/TVCG.2022.3175626	2016	cartograms	poor / excellent readability
10.1109/TVCG.2022.3175626	2017	scatter plots; 3D scatter plots	feature easy to perceive
10.1109/TVCG.2022.3175626	2017	flows, paths	able to read
10.1109/TVCG.2022.3175626	2018	flows; maps; paths	good visual design; ease of use
10.1109/TVCG.2022.3175626	2018	slides with multiple types of vis	easy to read; easy to follow
10.1109/TVCG.2022.3175626	2018	vector glyphs	easy / hard to interpret
10.1109/TVCG.2022.3175626	2018	multiple types; point-based; text-based	intuitive
10.1109/TVCG.2022.3175626	2019	calendar: mutliple charts in analytics view	easy to understand; fast to find the information
10.1109/TVCG.2022.3175626	2020	steamgraphs	good readability
10.1109/TVCG.2022.3175626	2020	multiple type of vis; surface-based; colormaps; glyph-based; volumes; continuous colors	readable; feature visible; recognizability; enable overview; unambiguous (obvious to understand); at a glance; able to determine; assessable; interpret
10.1109/TVCG.2022.3175626	2020	storyline; paths; timelines; node-links	see clearly
10.1109/TVCG.2022.3175626	2020	node-links; color-coded text	easy to understand
10.1109/TVCG.2022.3175626	2020	maps; VR	ease of use
10.1109/TVCG.2022.3175626	2021	bar charts; line charts	I am confused - it makes sense
10.1109/TVCG.2022.3175626	2021	infographics	readability
10.1109/TVCG.2022.3175626	2021	multiple; text-based; line-based; dot-based; maps; ...	easy to use for understanding
10.1109/TVCG.2022.3175626	2022	multiple charts; bar graphs; line charts; pie charts; dot based; areas	facilitate understanding
10.1109/TVCG.2022.3175626	2022	bar charts; line charts; VR	easy to learn; easy to understand
10.1109/TVCG.2022.3175626	2022	glyph based; surface based; bar charts	quick to read; easy to read
10.1109/TVCG.2022.3175626	2022	tables; text; path; node-link	interpretable
10.1109/TVCG.2022.3175626	2022	bar charts	clear to understand; easy to read

Table 5: 36 studies with user comments from which we extracted terms in Pool 2 (see Sect. 4.1.1).

DOI	Year	Type of visualization	Terms collected for item development
10.1109/INFVIS.2002.1173148	2002	node-links	readable; visible
10.1109/TVCG.2009.176	2009	labels; maps	legible
10.1109/VAST.2009.5332595	2009	timelines	readable; simple
10.1109/TVCG.2011.186	2011	maps; node-links	better; suitable; difficult; follow; lost
10.1109/TVCG.2011.160	2011	charts	decipher; confusing
10.1109/TVCG.2012.255	2012	node-links	rich; usable; comprehensibly; easily; simplicity; speed; understandable
10.1109/TVCG.2012.189	2012	paths	meaningful; view; simple; follow; understand; helpful
10.1109/TVCG.2013.151	2013	node-links	difficult
10.1109/TVCG.2013.191	2013	charts	readability; easy to see; easy to read; legible; ease of visibility; understand; clean; bold; untidy; cluttered
10.1109/TVCG.2013.233	2013	flowcharts	easy; see; difficult; clutter; use
10.1109/TVCG.2014.2346420	2014	node-links	occluded; confusing; recognize; simple; guide; intuitive
10.1109/TVCG.2015.2467872	2016	parallel coordinates plot	strong; easy
10.1109/TVCG.2017.2745941	2018	scatter plots; 3D scatter plots	judge; ease; clear
10.1109/TVCG.2017.2744118	2018	storylines	understand; recognize; disorientating
10.1109/TVCG.2018.2865192	2019	flows; maps; paths	easy; difficult; follow; unexpected; clear; distinguish; encoding; intuitive; sparse
10.1109/TVCG.2018.2864907	2019	motion patterns	overview; instantly; interpret; hard; cognitive
10.1109/TVCG.2019.2934784	2020	fact sheets	easy; understand; meaningful; present (?)
10.1109/TVCG.2019.2934557	2020	matrices	easy; perceive; helpful; distinguish; coding; obvious; effective
10.1109/TVCG.2019.2934337	2020	dot plots; motion paths; surface-based	discern; interpret; understand
10.1109/TVCG.2019.2934669	2020	line-based; bar charts	unclear; hard; meaning
10.1109/TVCG.2020.3030388	2021	multiple types: surface-based; colormaps; glyph-based; volumes; continuous colors	clear; intuitive; interpret; correctly; spot; easily
10.1109/TVCG.2020.3030358	2021	node-links; color-coded text	difficult; understand; easy; visualize; conceptualize
10.1109/TVCG.2013.76	2013	maps; areas	distracting; confusing; hard; lost; follow; chaos
10.1109/TVCG.2014.2329308	2015	maps; motion paths	easier; busy; messy; comfortable
10.1109/TVCG.2014.2337337	2015	node-links	cluttered; richer; neat
10.1109/TVCG.2018.2835485	2019	multiple types: point-based; text-based	intuitively; easy; hamper; useful; appropriate; confusing; distinguish; efficient; highlight; attract attention; recognize
10.1109/TVCG.2019.2906900	2020	node-links; paths	complex
10.1109/TVCG.2020.2968911	2021	glyphs	simple; complex
10.1109/TVCG.2021.3068337	2022	bar charts; ligne charts	distracting; messy; cluttered; disorganized; unappealing
10.1109/TVCG.2021.3114822	2022	multiple types: glyph based; timeline; line-based	useful; insight; easily; understand; prefer; useful; see; relevant; show; attention; obvious
10.1109/TVCG.2021.3114775	2022	animation of charts	balance
10.1109/TVCG.2022.3209477	2023	multiple types: node-links; matrix; dot based	understand; interpretation
10.1109/TVCG.2022.3209475	2023	bar charts; line charts; VR	occlusion; balance
10.1109/TVCG.2022.3209484	2023	networks	intuitive; easy, understand
10.1109/TVCG.2022.3209480	2023	flows; maps; paths	easy; hard; confusing; visible; see; crowded; convenient; show
10.1109/TVCG.2022.3144975	2023	tables; text; flow	intuitive; informative

Table 6: Terms related to readability retrieved from deductive and inductive methods (as described in Sect. 4.1)

unique terms	collected terms	overall counts	collected terms from studies	collected terms from comments	collected terms from survey
easi	easy	40	easy	easy	easy
understand	understand, understanding, understandable	36	understand, understanding	understandable, understand	understand, understanding
interpret	interpret, interpretation	16	interpret	interpret, interpretation	interpret
inform	information, informative	15	information	-	information, informative
clear	clearly, clear	15	clearly, clear	clear	clear, clearly
read	read, reading	14	read	read	reading, read
visual	visually, visual, visualization, visualize	12	visually, visual, visualization	visualize	visual, visually
easili	easily	12	-	easily	easily
clutter	cluttered, clutter	11	cluttered	cluttered, clutter	clutter, cluttered
data	data	11	data	-	data
readabl	readable, readability	11	readable, readability	readable, readability	readable
see	see	11	see	see	see
difficult	difficult	11	difficult	difficult	difficult
confus	confused, confusing	10	confused	confusing	confused, confusing
use	use, useful	9	use	use, useful	useful
hard	hard	8	hard	hard	hard
intuit	intuitive, intuitively	8	intuitive	intuitive, intuitively	intuitive
eas	ease	6	ease	ease	-
confid	confidence, confident	6	confidence	-	confident
overview	overview	6	overview	overview	overview
identifi	identify, identifiable	6	identify	-	identifiable, identify
show	show, shows	6	-	show	show, shows
mean	meaning, means	6	-	meaning	meaning, means
encod	encoding, encoded, encodings	5	-	encoding	encoding, encoded, encodings
visibl	visible, visibility	5	visible	visible, visibility	-
simpl	simple	5	simple	simple	-
relev	relevant	5	-	relevant	relevant
valu	value, values	4	value	-	values
meaning	meaningful	4	-	meaningful	meaningful
help	helpful, help	4	-	helpful	help
present	present, presentation, presented	4	-	present	presentation, presented, present
find	find	4	find	-	find
recogn	recognize	4	recognize	recognize	recognize
learn	learn, learned	4	learn	-	learned, learn
look	look	4	look	-	look
quick	quick, quickly	4	quick	-	quickly
attent	attention	3	-	attention	attention
complex	complex	3	-	complex	complex
answer	answer	3	answer	-	answer
effect	effective	3	effective	effective	-
follow	follow	3	follow	follow	-
distract	distracting, distracted	3	-	distracting	distracted
design	design	3	design	-	design
obvious	obvious	3	obvious	obvious	-
featur	feature, features	3	feature	-	features
arrang	arrangement	2	arrangement	-	arrangement
correct	correct, correctly	2	correct	correctly	-
text	text	2	text	-	text
compon	component, components	2	component	-	components
perceiv	perceive	2	perceive	perceive	-
glanc	glance	2	glance	-	glance
grasp	grasp	2	grasp	-	grasp
discern	discern	2	discern	discern	-
determin	determine	2	determine	-	determine
deciph	decipher	2	-	decipher	decipher
unexpect	unexpected	2	-	unexpected	unexpected
easier	easier	2	-	easier	easier
attract	attract, attracting	2	-	attract	attracting
insight	insight	2	-	insight	insight

C GENERATED ITEMS DETAILS

C.1 Table of 39 generated items

Table 7 contains all items generated as described in Sect. 4.2, and with additional details below.

C.2 A detailed example of how we built a candidate item

As mentioned in Sect. 4.2, to combine primary and secondary terms into candidate scale items, we referred to phrasing and conceptual patterns extracted from the statements collected in our expert survey (see Sect. 4.1.2). We extracted these patterns by analyzing syntactic roles and conceptual families of words in experts' statements.

C.2.1 Splitting complex proposed items

As a first step in this analysis, we split complex items (nested clauses and compound sentences) into individual statements, with a effort to put resulting single-clause sentences in a affirmative, active form. For example, as shown in Fig. 6, the sentence "*Viewing this chart, I understand the overall message that the visualization designer is trying to convey*" was separated in two statements:

1. "*Viewing this chart, I understand the overall message*", and
2. "*The visualization designer is trying to convey an overall message*".

In doing this, we transformed 128 proposal items into 147 statements. We then proceeded to the analysis of the structures and concepts in each statement. Here, we will present a detailed example of how we processed the following statement:

"I can recognize main characteristics of the data"

C.2.2 Categorizing words by role in sentences

As a second step, we identified the syntactic role and meaningfulness of each term regarding the statement:

→ (Subject) *viewer*

⟶ (Subject quality) *able*

→ (Verb) *recognize*

→ (Object) *data characteristics*

⟶ (Object quality) *main*

Our work here consisted in capturing the essence of what each statement described and making it consistent. For example, we transformed questions into affirmations, and discarded complements such as "in the visualization" or "on the screen". We also transformed some words at this stage. For example, we coded all instances of "I" our "you" terms into a generic "viewer" term, and all terms such as "chart", "visual data representation", etc. into a generic "visualization" term. We also transformed verbs such as "can" and "am able to" into a characteristic of the subject labeled "able" (as in the example above), which allowed us to focus on the visualization activity as the main predicate.

We refer the reader to our file of annotated statements in our supplemental material for a complete account of this process.

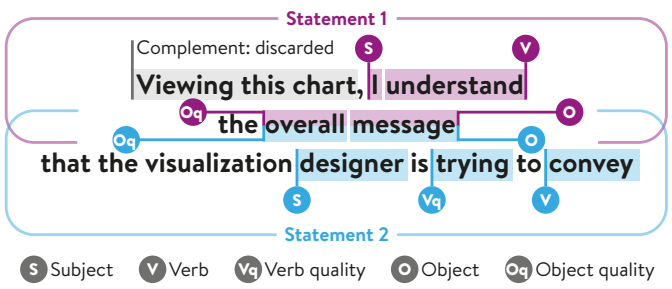

Fig. 6: An example of separation of clauses and syntactic role attribution, as described in Sect. C.2.

C.2.3 Coding conceptual families

For each type of role (Subject, Subject quality, Verb, etc.) in our annotated table of expert propositions, we extracted a table of all terms. Two researchers of the team independently attributed terms to abstract semantic groups: the CONCEPT families. They coded a main concept, and an additional one if needed. For example, the first author coded all following Object terms as DATA FEATURE: *trends, outliers, pattern, features, trend, patterns, clusters, groups, relationships, communities, activity, level, characteristics*. Then the two researchers discussed results harmonization on the main concept level, and the remaining uncertainties or unresolved conflicts were taken to the full research team for final decision. Only the main concept was saved for the next step of the work. We then mapped the CONCEPTS back to the original structure of each sentence. The example above thus became:

→ (S) READER

⟶ (Sq) SUPPORT

→ (V) QUERY

→ (O) DATA, DATA FEATURES

⟶ (Oq) KEY

With this approach, we produced an set of sentence patterns with 60 CONCEPTS from the original 176 unique terms collected in Pool 3 (see Sect. 4.1.2). 13 additional CONCEPTS were found from terms in statements that we did not include in Pool 3. Our syntactic analysis approach allowed us to synthetize the survey results in sentence patterns while keeping a refined association between original words and their aggregation as CONCEPTS. After having finalized the definition of CONCEPTS based on the survey content, the main author expanded the concepts attribution to terms from the other pools from the deductive method described in Sect. 4.1.1. It was not necessary to create new CONCEPTS at this stage.

This work served as guide for item writing: we referred to visual representations of expert statements patterns (for example, using Word Tree, Fig. 7), in combination with the dictionaries of term-CONCEPT associations.

These dictionaries (which we called "lexicons") were of particular importance because some words can have multiple meanings. For example "look" has different meanings between "it looks awesome" and "I don't know where to look"). What's more, at this stage all terms from our 3 pools had been reduced to their root (i. e., their *stem*). Stemming is useful to reduce the number of terms; however it also reduces semantic precision, for example when aggregating "information" "informative" in a unique "inform" *stem*. For these reasons, we decided it was important to refer to the conceptual categorization, which was built from words rather that from stems.

We refer the reader to our OSF Research log for a complete account of the writing procedure and a description of the files provided as supplemental material in the OSF repository.

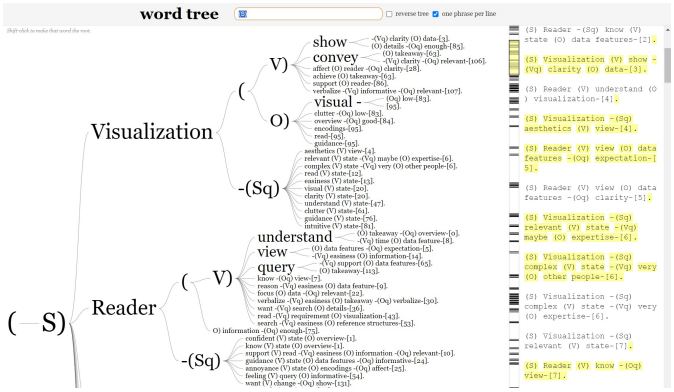

Fig. 7: Among other tools, we used Jason Davies's Word Tree online tool at `jasondavies.com/wordtree` to visualize conceptual patterns in collected expert propositions of statements for assessing perceived readability.

Table 8: Mean, median, and modes of experts ratings on a scale from 1 to 5 for each candidate item, as explained in Sect. 4.3

term	median	mean	largest mode	smallest mode	all modes	keep
answer	5	4.26	5	5	5	yes
clearEncod	5	4.23	5	5	5	yes
clearInform	4	4.03	4	4	4	yes
clutter	3	3.23	4	4	4	yes
complex	4	3.84	4	4	4	yes
confid	4	3.81	5	5	5	yes
confus	4	4.03	4	4	4	yes
correct	4	3.48	5	4	(4, 5)	yes
deciph	4	3.94	5	5	5	yes
determin	4	4.00	4	4	4	yes
discern	4	3.26	4	4	4	yes
distract	3	3.13	4	4	4	yes
effect	4	4.06	4	4	4	yes
encod	5	4.23	5	5	5	yes
find	4	3.81	5	5	5	yes
follow	4	3.81	4	4	4	yes
help	4	3.81	5	5	5	yes
identifi	4	3.90	5	4	(4, 5)	yes
inform1	4	4.29	4	4	4	yes
inform2	3	3.23	4	4	4	yes
interpret	4	4.23	5	5	5	yes
intuit	4	3.77	4	4	4	yes
makeSense	4	4.00	5	5	5	yes
meanCompon	4	4.16	4	4	4	yes
meanOveral	4	3.90	4	4	4	yes
obvious	4	4.23	5	4	(4, 5)	yes
organiz	4	3.58	4	4	4	yes
overview	3	3.58	3	3	3	yes
perceiv	4	3.97	4	4	4	yes
read	4	4.19	5	5	5	yes
readabl	4	3.90	4	4	4	yes
recogn	4	3.81	4	4	4	yes
see	4	3.90	4	4	4	yes
show	4	3.77	5	5	5	yes
simpl	4	3.97	5	4	(4, 5)	yes
understandEasi	4	4.06	5	5	5	yes
understandQuick	4	3.90	4	4	4	yes
valu	4	3.65	4	4	4	yes
visibl	4	3.65	4	4	4	yes

D Survey results for expert validation of items

We provide in Table 8 the expert ratings on our 39 candidate items, which we collected in the study described in Sect. 4.3. All candidate items received a mean, mode and median above 3. Therefore we kept them all for the scale development process, starting with the pre-test study we describe next in Appx. E.

We also provide the questionnaire used for this study and the collected data in our supplemental materials on OSF. Experts had the opportunity to comment: we read and integrated those comments in our reflection process. Three comments pointed out that the statement "I find parts of the visualization distracting" was lacking the "[reversed item]" tag that we put for others statements where a high agreement would indicate lower readability (e. g., "I find this visualization complex to read").

E Pre-testing study details

In this section of the appendix, we provide important details on the pre-testing study we conducted before administrating the exploratory survey on Prolific for the PREVis development phase in Sect. 5.

Visualization stimuli. In this survey we used 4 stimuli from the VLAT [59] (line chart, histogram, choropleth map, and bubble chart) and 2 of their associated VLAT questions, as found on the reVISit webpage [100]. This choice was motivated by the need to test our rating items among a range of low to high readability situations. In the absence of an existing set of readable to non-readable visualizations, we considered two main factors: efficiency of visual design, or difficulty of reading tasks. In this pre-test, we chose the latter because it offers a principled assessment of the difficulty of design-task pairs. We provide additional details about the choice of stimuli and associated tasks in our OSF Research log.

Participants. We invited participants by sending an email to the staff and students of the first author's university. We did not compensate participants for taking part in this study. We received 12 positive answers but one participant canceled their interview slot. As a result, 11 people participated in one of this study's 3 rounds (4 men, 7 women; age: mean 30, min: 22, max: 47; highest degree completed: 5 Bachelor, 6 Master).

Procedure. We conducted interviews online [90] or in-person, depending on the participants' preferences. In both case, participants connected to our online survey platform. We first asked them to read and agree to the informed consent form, and to answer background questions about their gender, education, and English fluency. Participants then completed a training think-aloud task. When they agreed to continue the study after this task, we started the recording of their voice and screen. We then randomly assigned the participants to one of 4 possible stimuli. In round 1 and 2, we ensured that each stimulus was seen at least by one person. Round 3 had only participant and the stimuli was assigned at random. Participants answered 2 reading tasks and one comprehension check about the stimulus, after what they rated the visualization using the 39 candidate items and a 7-point agreement scale. The interviewer observed the participant's behavior to identify hesitations, and occasionally prompted them to elucidate problems if the participants had gone silent. Interviews duration ranged from 45 to 93 minutes.

Data analysis. We used the problem respondent matrix [25] to analyse the content of the interviews. This tool allows to break down the questionnaire answering process into 3 stages: understanding (i. e., making sense of the question), task performance (e. g., forming an opinion, or recalling an experience), and response formatting (e. g., choosing a Likert-scale answer option); and to categorize respondents potential issues into 5 groups of problems: lexical, temporal, logical, computational, and omission/inclusion. Cognitive interviews are qualitative methods to detect problems, the goal is not to perform quantitative analysis. We did not conduct analysis of other data such as comments on the definitions of "reading", "understanding" and "interpreting" or reading tasks.

Table 9: Our respondent problem matrix for coding interviews.

PROBLEM TYPE	RESPONSE STAGE		
	Understanding	**Task performance**	**Response Formatting**
Lexical	ULe	TLe	RLe
Logical	ULo	TLo	RLo
Computational	UCo	TCo	RCo
Omission/Inclusion	UOm	TOm	ROm

Items changes. Over the 3 rounds of interviews we rephrased items that lacked clarity, and dropped items that could not be clarified. We present the final list of items statements in Table 12. In our supplemental material, we provide a summary of item changes over the 3 rounds. We also share the codings from cognitive interviews in round 1, round 2, and round 3. Finally, we share our methodological notes on cognitive interviewing as separate material [1].

In particular, we noticed that participants often incorrectly rated reversed items (e. g., "I find parts of this visualization distracting"). In many cases, they did not notice it until asked to talk more about their choice of answer. We found that, most items being of positive valence (e. g., "I find this visualization easy to read"), participants tended to associate the right side of the Likert options (towards "Strongly agree") to having a "good opinion" of the visualization regarding the thematic (e. g., understanding or distraction), rather than an agreement with the statement itself. Therefore, after round 3 we reworded items of negative valence with a negative turn of phrase (e. g., 'I don't find distracting parts in this visualization").

Survey design changes. Comments from participants and observations from the interviewer also allowed us to refine our survey's design. As such, in the first round of our pre-test study, participants answered 39 rating items and one attention check on a single screen (see Fig. 8). We imagined it would make the task less tedious, but we found that respondents could easily miss an item in this setting. When it was the case, participants then had to scroll back up to spot the empty item, which they found tedious. As a result, we designed the final survey to display each rating item on an individual screen Fig. 9, and we added an option to use the "Enter" key from the keyboard to access the next screen more quickly than with a mouse interaction. We also refined the presentation of the 7 points Likert-scale over rounds: in the first round, respondents saw numbers from 1 to 7; while in the final survey, they could choose between 7 points, all labeled (from "Strongly disagree" to "Strongly agree"). We also added a separate option labeled "I don't know (please elaborate)" with a short text field. Although such an answer option generates missing data in the collected answers, it can also reduce noise and is generally recommended in surveys [22].

We provide a visual summary of changes in survey design during and after pre-test as supplemental material.

An other important outcome of the pre-test study was our decision not to use the VLAT items as stimuli for our final exploratory survey. This was motivated by our observation that the stimuli were visually very clean, and participants tended to comment a lot about their *interpretability* rather than *readability* in the think-aloud procedure. We thus decided to focus the design of our final stimuli on different criterion, as we describe next in Appx. F. We detail this decision a little more in our OSF Research log.

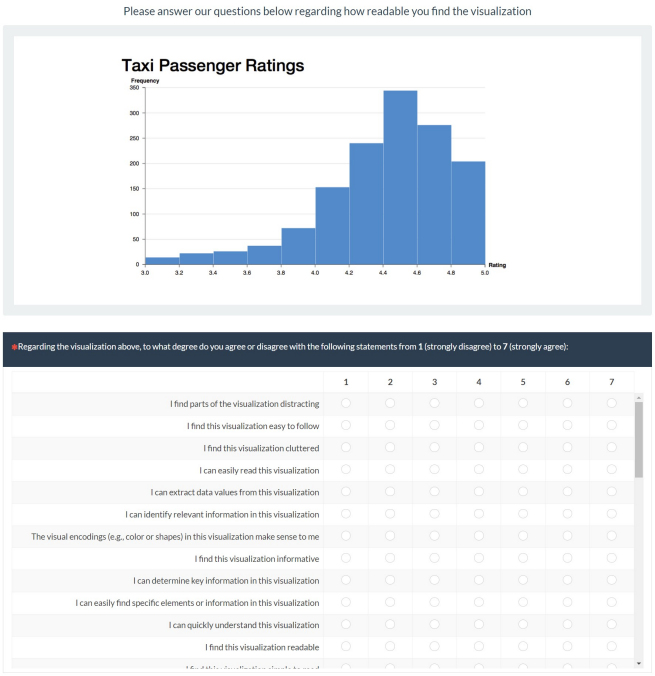

Fig. 8: Presentation of a rating item in round 1 of the pre-test study.

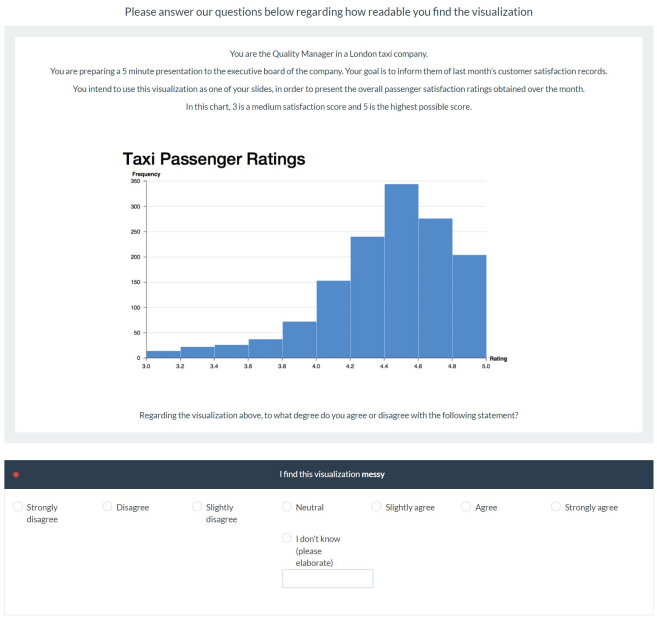

Fig. 9: Presentation of a rating item in round 3 of the pre-test study.

[1] We do not share this document in the main supplemental material OSF repository as it contains images from previous work which cannot be redistributed under the CC BY licence

Table 7: Outcome of the item generation step (Sect. 4.2): 39 items generated from primary and secondary relevant terms.

Item statement	**Primary term**	**Auxiliary term(s)**
I can answer questions about the data after reading this visualization	answer	data, read
The visual encodings (e.g., color, shapes...) in this visualization make the information clear to me	clear	encod, inform
This visualization shows information in a clear way for me	clear	inform
I find this visualization cluttered [reverted item]	clutter	-
I find this visualization complex to read [reverted item]	complex	read
I am confident in my interpretation of this visualization	confid	interpret
I find this visualization confusing [reverted item]	confus	-
This visualization allows me to correctly interpret the data	correct	data
I find the visual encodings (e.g., color, shapes...) difficult to decipher in this visualization [reverted item]	deciph	difficult, encod
I can determine key information in this visualization	determin	key, inform
This visualization allows me to discern elements of the data	discern	data
I find parts of the visualization distracting [reverted item]	distract	-
This visualization effectively shows the data to me	effect	data, show
I can understand how the data is encoded in this visualization	encod	data, understand
I can easily find specific elements or information in this visualization	find	easili, inform
I find this visualization easy to follow	follow	easi
This visualization helps me understand the data	help	understand, data
I can identify relevant information in this visualization	identifi	relev, inform
I can easily retrieve information from this visualization	inform	easili
I find this visualization informative	inform	-
I find this visualization easy to interpret	interpret	easi
I find this visual design intuitive	intuit	visual, design
The visual encodings (e.g., color, shapes) in this visualization make sense to me	make sense	encod *(Note: "make sense" is used as a replacement for "meaningful")*
I can easily interpret the overall meaning of the data visualization	mean	overall, interpret
I can understand what the visual components of the visualization mean	mean	understand, visual, compon
It is obvious for me how to read this visualization	obvious	read
I find this visualization well organized	organiz	*(Note: "organiz" is used as a replacement for "arrang")*
This visualization provides me with a good overview of the data	overview	data
I can easily perceive data features (e.g., trends, minimums, outliers...) in this visualization	perceiv	easili, data, featur
I can easily read this visualization	read	easili
I find this visualization readable	readabl	-
I can recognize data features (e.g., trends, minimums, outliers...) in this visualization	recogn	data, featur
I can clearly see data features (e.g., trends, minimums, outliers...) in this visualization	see	data, featur, clear
This visualization shows the data in an appropriate manner for me	show	data
I find this visualization simple to read	simpl	read
I can easily understand this visualization	understand	easili
I can quickly understand this visualization	understand	quick
I can extract data values from this visualization	valu	data
I find data features (e.g., trends, minimums, outliers...) visible in this visualization	visibl	featur

F Exploratory survey stimuli visualizations and tasks details

This section complements Sect. 5.1, as it provides additional details regarding the content of the survey used to develop PREVis.

We provide a summary of the stimuli and their reading tasks as supplemental material.

F.1 6 stimuli visualizations

We wanted to develop and validate an instrument that could be used in a variety of visualization and readability situations. To that end, for our exploratory data collection we aimed at creating a sample of visualizations stimuli to test with variety in terms of:

- **underlying data** w.r.t. data domain (what they represent), data structure (e. g., tables and networks), number of data points, and number of data attributes;
- **visualization idioms** used (e. g., glyph-based, bar chart, point-based, line-based, node-link, areas, surfaces and volumes, matrix, text-based, continuous color. . .); and
- **presumed readability** for the selected tasks and target audience.

It would not have been feasible to engage in a systematic exploration of all possible variations for so many characteristics; however we attempted to obtain at least some variability for all of them.

Table 11 summarizes the design characteristics of the visualizations we used as stimuli in our crowd-sourced experiment (for which the procedure can be found in Sect. 5.3). Stimuli A B C (shown in Fig. 10) had presumably rather good readability, while D E F (also in Fig. 10) had lower presumed readability based on: difficult encodings in D, unfamiliarity for E, and visual clutter in F.

F.1.1 Number of visualization stimuli—Pilot study

We had a limited Prolific budget which would only allow us to fund 156 hours worth of participation. We thus ran a pilot study to assess how much time was needed for respondents to answer our rating items. The results allowed us to calculate how many visualizations we would be able to test in the exploratory survey within our budget restrictions.

We recruited 14 Masters students from a visualization class we teach. Participation was voluntary and they did not receive any compensation. After agreeing to a consent form and answering 5 demographic questions (color vision deficiency, age, gender, english fluency, and education), each participant saw 2 visualizations out of a set of 4 possible visualizations and answered all questions of: 2 reading tasks, 1 comprehension check and 29 candidate rating items, and 1 attention check in the form of a rating item.

The mean time required to complete the survey was 20.19 min (std = 3.8). Our results (see Table 10) indicated that Masters students needed about 2 minutes to answer the consent form and demographic questions, leaving 18 min for 2 stimuli. We estimated that crowd-sourced workers from Prolific would be a little faster as they tend to minimize the time spent on tasks. In addition, we would retrieve demographic information from Prolific data (except color vision deficiency). We thus estimated each Prolific participant would need 7–8 minutes to complete our survey with one visualization.

As fatigue has been documented to appear after 10 minutes in crowd-sourced studies [113], our goal was to allow respondents to complete the survey under this threshold. This fortified our decision to show only one stimulus to each participant, rendering our study design effectively cross-sectional with independent groups assessing each visualization.

Our Prolific budget was 156 hours worth of participation: in the best case scenario of 7 minute for each visualization, we could reach a maximum of 1330 ppeople—1170 for 8 minutes. To reach our target sample size of 300 participants per visualization (more details on that in our OSF Research log), we would be able to use 4 different stimuli. We later received more funds and could add 2 further stimuli and expand on our stimulus characteristics.

As a result, we designed 6 stimuli for our exploratory survey; and we used stimuli A B C D for our first run of the survey; only later did we run a separate survey with survey E and F.

Table 10: Time spent by participants in our pilot study. Each participant randomly saw 2 our of 4 visualizations.

Participant	Total time spent on survey (min)	Time spent on consent and demopgraphics (min)
537729235	18.9	1.6
2097021850	20.7	2.4
2130749281	19.4	4.7
190639113	27.5	1.8
442561266	18.6	2.4
733260304	17.1	2.0
575889510	19.8	1.1
805650123	25.8	1.3
406801568	14.4	1.0
925228469	23.9	1.6
1655604134	22.6	2.7
1020624217	19.0	1.5
293208060	14.4	1.4
383121506	20.7	2.2
mean	**20.2**	**2.0**
std	**3.8**	**0.9**

In the following subsections, we briefly address each of the characteristics cited above and how we considered it in our stimuli design.

F.1.2 Underlying data

Here we briefly describe our rationales regarding the data we plotted in our stimuli visualizations. The first concern we had was that our instrument was not meant to capture a participant's knowledge of the domain. Although this factor could influence a reader's ability to correctly interpret a visualization, it is unclear whether it is useful to measure at the *reading* level. Therefore, a fundamental requirement for our stimuli was that we did not want readers to require any domain knowledge regarding the represented data. Keeping this in mind, we explored possible data characteristics for which we should consider variations in our stimuli to best capture the range of possible factors affecting readability during the exploratory study.

A first point of attention was the nature of *what* the data represented. In their seminal work, Card *et al.* [20] distinguished **physical data** with a spatial mapping—"*the human body, earth, modecules or others*"—from **nonphysical information**—"*such as financial data, business information, collections of documents, and abstract conceptions*". In our survey, stimulus C was a spatial representation (world map) while others were not— namely, we represented financial data (prices in B, budget in D, profits and losses in F), genealogical information in E, and internet bandwidth speed in A.

A close concept to that of *what* the data represents is the notion of *how* the data is structured: different structures might entail different readability factors in their representations. Munzner [70] identifies four **dataset types**: tables, networks, fields and geometry (spatial). Stimuli A, B, D F are represented from tables, while E is a representation of a network, and C is a spatial representation.

Another major factor influencing design choices in data visualization is the number and type of **data attributes** to represent. The data we used for all stimuli has categorical and numerical attributes. E is somewhat different in the sense that it produces countable elements in the visualization from entirely categorical data.

Lastly, the number of represented **data points** can create visual clutter and make it harder for a reader to retrieve individual values; and, in dense layouts, large amounts of data points can cause overplotting, impeding the visual detection of higher-order patterns in the data. We did not provide extreme cases of data points density as to not overwhelm our participants, but the represented data in our visualizations ranged from 7 bars in A or 3 lines in B, to 36 lines in F. In C, we plotted 176 countries and our dataset contained data values for 122 of them; but we labeled only 10.

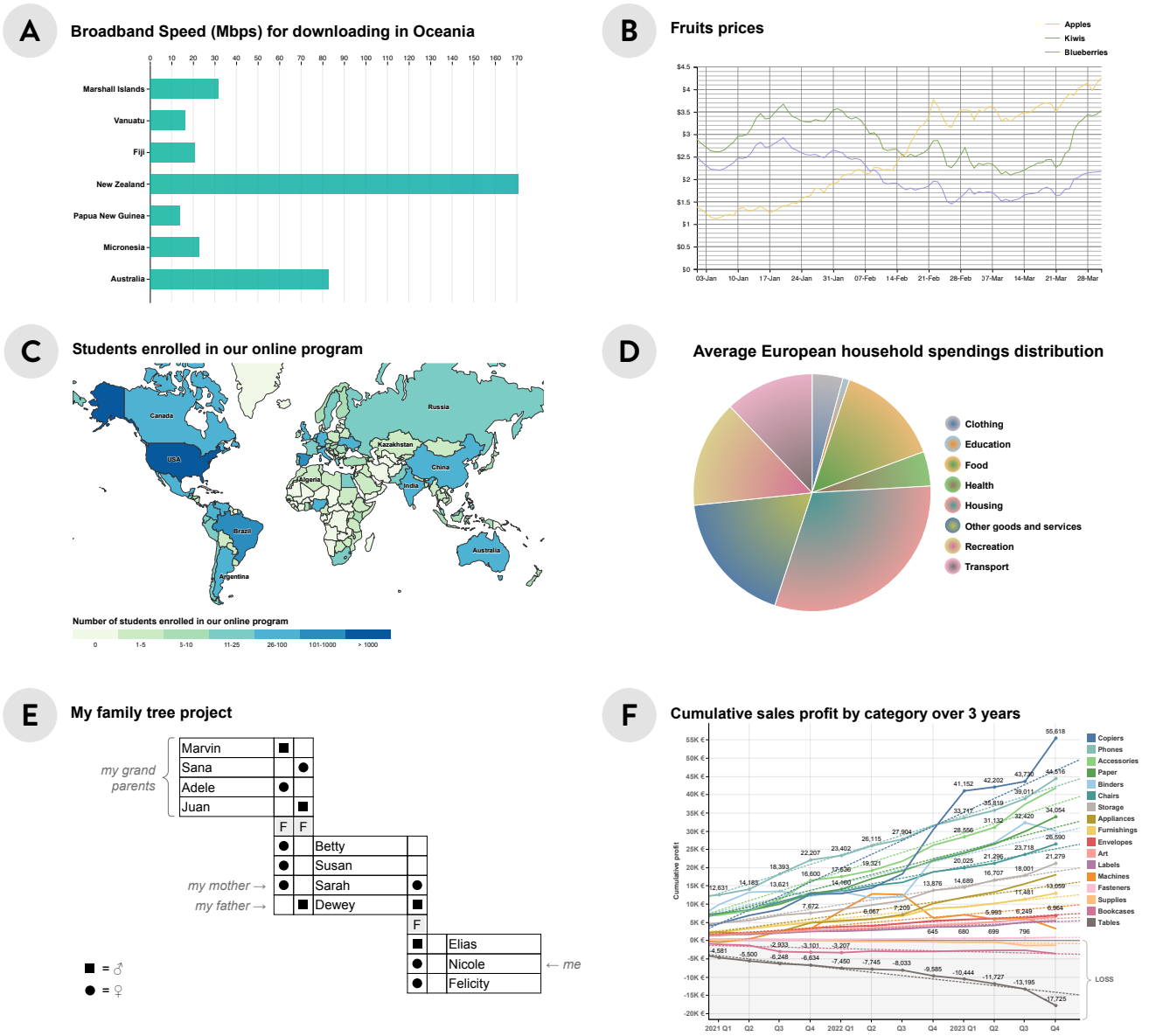

Fig. 10: The 6 stimuli visualizations we created for our exploratory survey to develop PREVis.

Table 11: Design characteristics of the 6 stimuli used in our survey and how they related to the 3 families of visualization characteristics possibly influencing readability.

Stimuli	**Display: encodings, data attributes** (on a canvas of max 800 x 500px)	**Individual: idiom familiarity**	**Task difficulty** in VLAT [59]
A - Bar chart	2 visual variables, few entities (7 bars), clear layout	Very familiar or intuitive	Easy
B - Line chart	3 visual variables, few entities (3 lines), cluttered layout (grid)	Very familiar or intuitive	Easy
C - Choropleth map	4 visual variables, many entities (world countries), simple encodings	Very familiar or intuitive	Easy
D - Pie chart	2 visual variables, few entities (8 categories), messy encodings (colors, labels,)	Very familiar or intuitive	Easy
E - GeneaQuilts (family tree in a matrix style [11])	3 visual variables, reasonable amount of entities (11 people), simple encodings	Not familiar, possible to infer mapping rules from labels with effort	- (not in VLAT)
F - Many lines chart	4 visual variables, many entities (18 categories x 2 types of lines), overplotting (too much information)	Somewhat familiar intuitive (people might no be familiar with general trend dashed lines)	Easy

F.1.3 Visualization idioms

Our preliminary work led us to think that readability could be affected by readers' familiarity with a particular idiom [58] or their visualization literacy [6, 13, 59]. Therefore, we needed our exploratory data to also encompass this dimensions. That being said, we did not aim at creating an instrument dedicated to measuring such a skill in participants. Therefore we used five idioms that were widely familiar to broad audiences, and one unfamiliar type of representation.

For familiar idioms we used: bar chart in A, line chart in B and F, choropleth map in C, and pie chart in D. It's worth noting that, while the line chart has widespread use as we use it in B, we also plot trend lines in F, possibly adding to the reading difficulty for this last stimulus.

For the unfamiliar visualization, we considered Parallel Coordinate Plots (PCP), which Lee *et al.* used in the VLAT [59] validation study. However, it can be very difficult for readers to intuitively guess how to read a PCP—if not impossible. In fact, Lee *et al.* provided participants with a training in their study design. We, on the other hand, would not provide a training for other stimuli. Deviating from our study parameters for one of our 6 independent groups would endanger the reliability of our findings.

Instead, we chose to use a visualization idiom that is both novel to our audience and "self-teachable": the GeneaQuilts technique [11], a matrix-based representation of genealogical data. GeneaQuilts are most useful to represent large genealogy datasets; but with small dataset, they provide a very interesting combination of qualities for our current work: they can look cryptic at first sight, but their design is minimalist. What's more, given basic knowledge of the family links in the represented genealogy, a viewer can autonomously deduce how the GeneaQuilts encodings work (see the Simpson's family GeneaQuilts representation in Fig. 11). We implement GeneaQuilts in E.

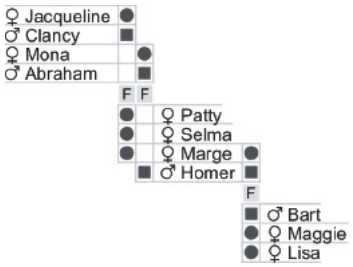

Fig. 11: GeneaQuilts Visualization of the Simpson Family in Bezerianos *et al.*'s original work [11]. Image is © 2010 IEEE, used with permission.

F.1.4 Visual display design choices to alter readability

For each visualization we created design variations that presumably would affect readability for respondents. We summarize our choices in the following paragraph a summary. Details can be found in a separate document from our supplemental material.

Firstly, our goal was to produce an instrument that would be usable in wide variety of ***likely*** visualization reading situations. Therefore, we did not include completely chaotic encodings (such as a complete mismatch of datatypes with idioms like plotting time trends on a pie chart). The furthest we went into that direction was the use of gradient colors encodings in D without in-chart labels. Secondly, our goal was to avoid potential influence of aesthetics judgement—for which there is already an existing measuring scale: [46]—on perceived readability. therefore, we attempted to keep a relatively clean an professional look for all images we produced.

To introduce design choices that would presumably increase or decrease perceived readability, we introduced facilitators (e. g., good separability of colors or direct labeling) and difficulties (e. g., absence of data label coupled to an arbitrary sorting of the accompanying legend).

F.1.5 Color impairment simulations

For stimuli relying on color encodings for data attributes, we chose color that would remain distinguishable for people with color vision deficiencies. We assessed the results in Adobe Illustrator using the View > Proof Setup for deuteranopia and protanopia, as shown in Fig. 12.

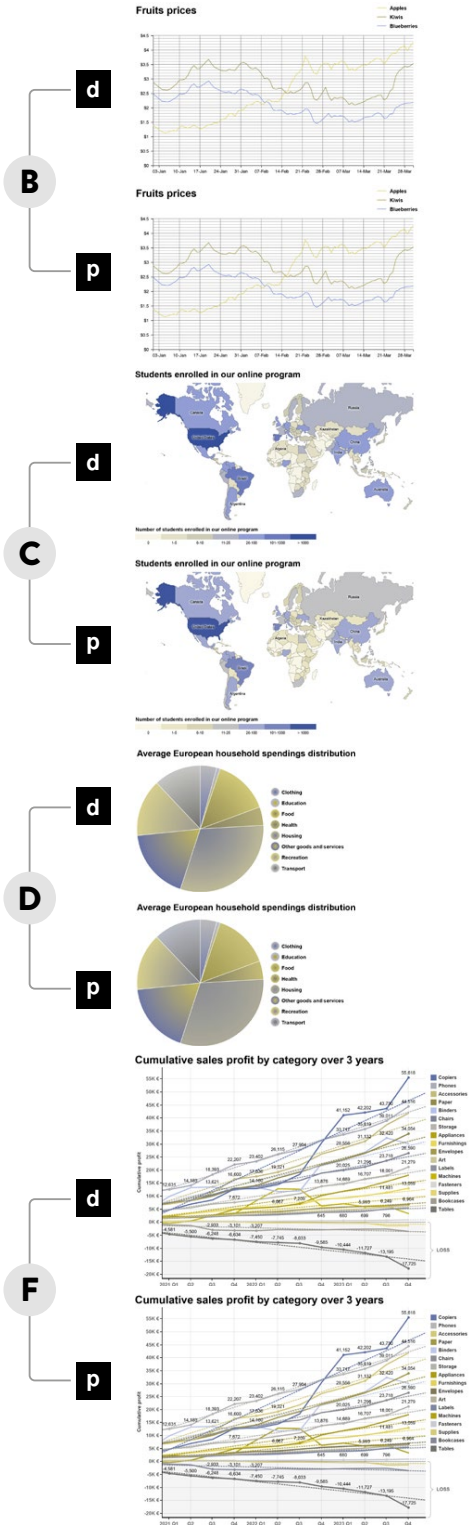

Fig. 12: Deuteranipa (d) and protanopia (p) simulations in Adobe Illustrator for all stimuli relying on color encodings for data attributes in our exploratory survey.

F.2 Reading tasks

Before asking participants to assess their perception of readability in a visualization, we need them to read it. Since this reading experience will shape their opinion on readability, we considered them as an

integral aspect of our stimuli for this exploratory study. Therefore, we paid a particular attention in selecting select tasks that were in the scope of "reading." We provide details on our review of three task taxonomies [3, 15, 28] and how we also referred to item difficulty from VLAT items [59] to form our starting set of reading tasks before selecting stimuli-task pairs.

In contrast with our difficulty-based approach during pre-test (see Appx. E), in our final survey we chose tasks that demonstrated low discriminating power in VLAT and were easy to perform, meaning that they did not require specific levels of visualization reading skills in our target audience. This could somewhat reduce the ability of our instrument to capture information about the respondent's visualization literacy; but then again, our goal was to focus on multiple factors of readability and we had already accounted for the influence of individual knowledge on answers variance by implementing the unfamiliar GeneaQuilts visualization in E.

F.2.1 Rating items

Table 12 presents the list of rating items we presented to participants after they answered reading questions. This list was refined from the initial 39 candidate statements through the pre-test study described in Appx. E.

F.3 Two rounds of survey implementation

We ran the survey in two separate rounds: the first one over the course of one week with stimuli A B C D, and the second with stimuli E and F, over the course of 1 + 2 days, separated by a 2 week delay for transferring additional budget to the Prolific platform. We did not make any change between the two rounds other than adding E F—and removing A B C D for which we had already collected our exploratory data in the previous round. We share printouts of the first round's survey and the second round's survey as supplemental material.

We pre-registered (`osf.io/4dcav`) our first round of the study. We updated the pre-registration on on March 9, 2024, after receiving the additional funds and before running our second round on Prolific; however, we encountered a glitch with the update feature, which we only noticed later. We contacted the OSF team and they manually fixed the registration on March 18, 2024.

Fig. 13: An example screenshot from our exploratory survey, consisting of: (1) a stimulus visualization (D: pie chart) with a title and a short contextual explanation, (2) a candidate rating item ("find") where keywords have been highlighted, and (3) 8 answer options consisting of a 7 points Likert-scale with individual labels, and one "I don't know / Not applicable" option, allowing the participant to elaborate on their reason for choosing this answer if they wish.

Table 12: 29 items and one attention check presented to participants in the exploratory survey to rate the readability of visualization stimuli.

Item code	Item statement
answer	I can easily **answer some questions** about the represented data with this visualization
clearData	This visualization shows the data in a **clear way** for me
clearRepresent	The representation of the data makes the **information clear** to me in this visualization
complex	I don't find this visualization **complex to read**
confid	I am **confident** in my understanding of this visualization
confus	I **don't** find this visualization **confusing**
crowd	I **don't** find this visualization **crowded**
deciph	I **don't** find the presentation of the data **difficult to decipher** in this visualization
distinguish	I can easily **distinguish individual elements** of the represented data (for example individual lines, or dots, or areas, or colors...)
distract	I **don't** find parts of the visualization **distracting**
effect	This visualization **effectively** shows the data to me
find	I can easily **find specific elements** in this visualization
identifi	I can easily **identify relevant information** in this visualization
inform	I can easily **retrieve information** from this visualization
lost	I **don't** feel **lost** trying to read this visualization
meanElem	I can easily understand what the **different elements** of the visualization **mean**
meanOveral	I can easily understand the **overall meaning** of this data visualization
messi	I **don't** find this visualization **messy**
obvious	It is **obvious** for me how to read this visualization
organiz	I find this visualization well **organized**
read	I can **easily read** this visualization
readabl	I find this visualization **readable**
represent	I can easily understand **how the data is represented** in this visualization
see	I can **clearly see** data features (for example, a minimum, or an outlier, or a trend) in this visualization
simpl	I find this visualization **simple to read**
understandEasi	I can **easily understand** this visualization
understandQuick	I can **quickly understand** this visualization
valu	I can **read data values** from this visualization
visibl	I find data features (for example, a minimum, or an outlier, or a trend) **visible** in this visualization
attentionCheck	For calibration purposes, please select **slightly agree** with this item

G Missing data handling for exploratory survey

As we provided an option to answer "I don't know" to rating questions in the exploratory survey, there was missing data in our collected ratings. Missing data, if not carefully handled, can impede the reliability of data analysis results. In particular, missing data can affect factors extraction in EFA [67]. Mirzaei *et al.* [69] provide a decision tree on handling missing data in surveys. We implemented the two checks they recommend in our data analysis code in `R`:

Step 1: Calculate the percentage of missing data. We calculated the amount of missing data and checked it again the thresholds Mirzaei *et al.* proposed:

- $<$ **5% missing data**: missing data is negligible and researchers may choose to handle missing data with deletion or imputation methods without significantly affecting their subsequent analysis;
- **5-10% missing data**: a grey area, where Mirzaei *et al.* recommend that the researcher refer to the theory regarding the phenomenon of interest before deciding on the next step;
- **10-40% missing data**: missing data is not negligible, but researchers may use imputation methods, depending on the results of the second step;
- $>$ **40% missing data**: missing data is too high for imputation methods: researcher should conduct a quantitative and qualitative investigation.

Step 2: Perform Little's test of missingness. A significant p-value result for this test [61] indicates that the null hypothesis that data is Missing Completely At Random (MCAR) is rejected, and therefore that a pattern exists to the missing data. When data is not MCAR,

We report the results of the missing data analysis in Table 13.

Table 13: Survey-wise and stimulus-wise frequencies of missing data and p-value in Little's test of Missing Completely At Random (MCAR) [61].

Dataset	Amount of missing data	Little's test p-value
Full survey	0.25%	<0.001 (data is MCAR)
Stimulus A	0.25%	0.624 (data is not MCAR)
Stimulus B	0.14%	<0.001 (data is MCAR)
Stimulus C	0.22%	<0.001 (data is MCAR)
Stimulus D	0.45%	0.109 (data is not MCAR)
Stimulus E	0.84%	0.013 (data is MCAR)
Stimulus F	0.28%	0.268 (data is not MCAR)

The tests on the complete dataset showed that all data was 0.25% and MCAR. Each individual stimulus' dataset also had less that 5% of missing data; however, Little's test of missingness showed that data was not MCAR in the datasets from stimuli A, D and F were not MCAR. As a result, we decided to use imputation methods, which tend to perform better than deletion methods according to the literature [67, 69].

We used the `mifa` package in `R` to generate the correlation matrix on which the EFA would be based. This package was developed specifically to perform multiple imputation for EFA [71] and relies on the `mice` package to perform multiple imputation Mutlivariate Imputation by Chained Equation for the estimation of a covariance matrix of incomplete data.

As participants were able to comment on the reason why they chose the "I don't know / Not applicable" answer option, we also report all comments collected this way along with the number of time respondents chose this option (regardless of whether or not they commented on it) as supplemental material.

H Correlation matrices from exploratory survey data

Here, we present the correlation matrix for the full dataset (Fig. 14), and then individual correlation matrices for each of our 6 stimuli in the survey (Fig. 15 to Fig. 20). We can make a few preliminary observations on these matrices:

- correlations are all positive and $>$ .3, which is considered moderate. Most correlations are $>$ .5, which is good, ans some $>$ .7, which is considered strong.
- correlations seem overall lower for situations that we expected to be of "higher readability" (stimuli A, B and C). It might be related to a ceiling effect from the Likert scales, where people chose the maximum ratings more often. As a result, relational structures are less visible in these matrices. We see structure more clearly in "bad" readability conditions than in good ones. Studying readability might be better achieved with "bas readability" than with good ones?
- although these matrices are organized in alphabetical order (on purpose so that the order remains stable to compare multiple analyses), making it more difficult to observe correlation groupings, we can see similar pattern
- a few groups of items always highly correlate across stimuli, which means we might expect to find them grouped in factor analyses with more than one factor:
  - `clearData + clearRepresent`
  - `read + readabl`
  - `understandQuick + understandEasi (+ simpl)`
  - `see + visibl`, which also generally do not correlate with other items. Therefore we expect these two form a single factor, unrelated to other items.

I Appropriateness of data for EFA

Before conducting the analysis, we needed to confirm whether our data was suitable for EFA. Following recommendations We assessed the univariate normality with a Shapiro Wilk's test, and multivariate normality with a Mardia test. Both tests showed significant results, indicating that our data violated the normality assumption. In such cases it is recommended to use a Principal Axis (PA) factoring method, which do not entail distributional assumptions [34]. We then tested the factorability of our correlation matrix using Bartlett's test of sphericity, as well as the Kaiser Meyer-Olkin (KMO) test, following reference work [99]. Bartlett's test yielded a p-value of 0, and all individual items' KMO values were above 0.7. Based on these results, we confirmed that our data's correlation matrix was factorable using a Principal Axis factoring method.

J Exploratory Factor Analysis: scree plots

Here, we provide all scree plots generated as part of our EFA described in Sect. 5.5.1. Scree plots show how much of the data can be explained with 1 to N factors, N being the number of measured variables (in our dataset: 29 items), and using *eigenvalues* of factors as values on the y axis. In the context of EFA, an eigenvalue represents the amount of signal (i. e., information) captured by a factor [30]. We plot on the same graph the results of parallel analyses, which show how much variance the same number of factors would capture for a randomly generated dataset of the same size as ours.

Visual analysis of such scree plots is two-fold:

- **Identifying elbows:** the elbow is the point in the slope after which a line begins to level off, meaning that adding new factors does not explain considerably more variance in the data. It is often used as a heuristic for determining the number of factors to retain.
- **Identifying a crossing point between lines:** this is the visual representation of a parallel analysis. We can observe when the line plotted from EFA on our dataset crosses the line from similiar EFA on a randomly generated dataset of same size. The logic to this approach is that the eigenvalue of the last retained factor should exceed that of an eigenvalue from random data.

In this appendix, we first present the scree plot for the full dataset (Fig. 21) on which we based on main analysis; and then individual scree plots for each of our 6 stimuli in the survey (Fig. 22 to Fig. 27), which we examined to confirm that the slopes exhibited similar characteristics across stimuli.

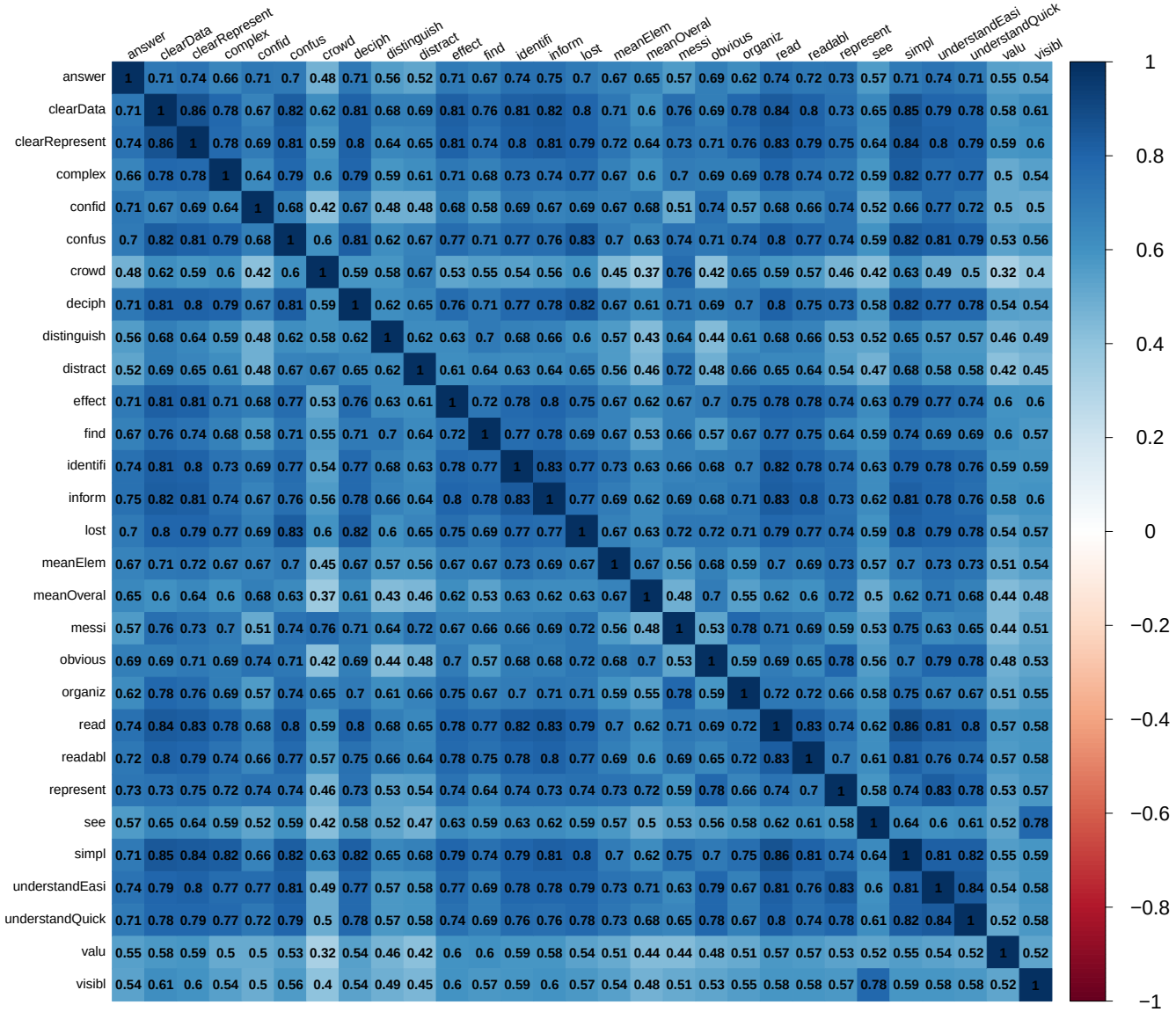

Fig. 14: Correlation matrix from our survey data after missing data treatment detailed in Appx. G. All items have positive correlations, meaning that they tend to co-vary in the same direction. The correlation value indicate to what extent: in here, we observe that all items have a correlation >0.3 with all other items, and a large majority exhibit correlations >0.5. The correlation matrix serves as input data for Exploratory Factor Analysis.

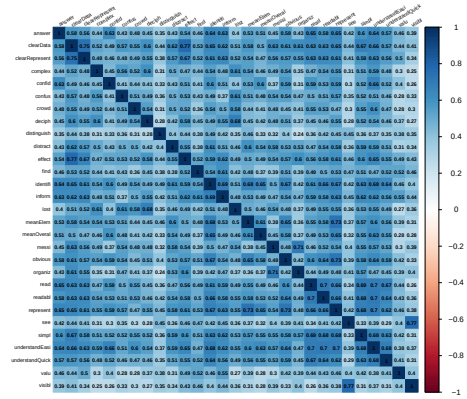

Fig. 15: Correlation matrix from our survey data for **stimulus A** (bar chart) after missing data treatement detailed in Appx. G.

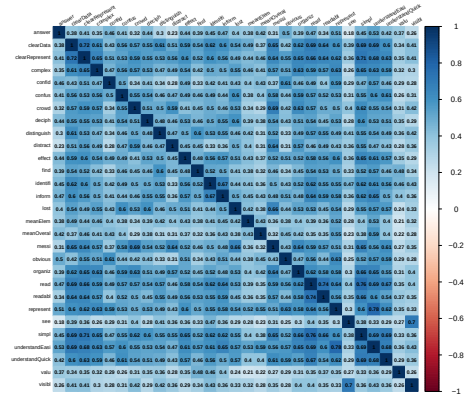

Fig. 16: Correlation matrix from our survey data for **stimulus B** (line chart) after missing data treatement detailed in Appx. G.

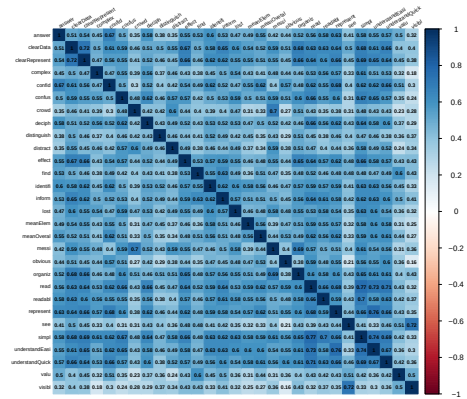

Fig. 17: Correlation matrix from our survey data for **stimulus C** (choropleth map) after missing data treatement detailed in Appx. G.

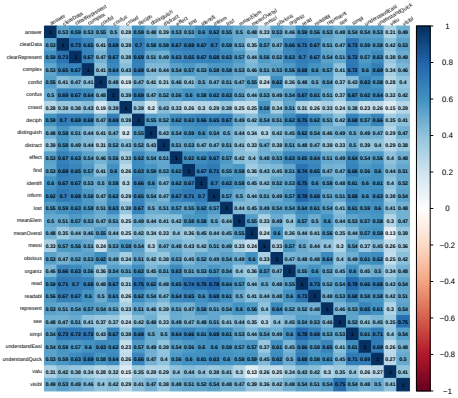

Fig. 18: Correlation matrix from our survey data for **stimulus D** (pie chart) after missing data treatement detailed in Appx. G.

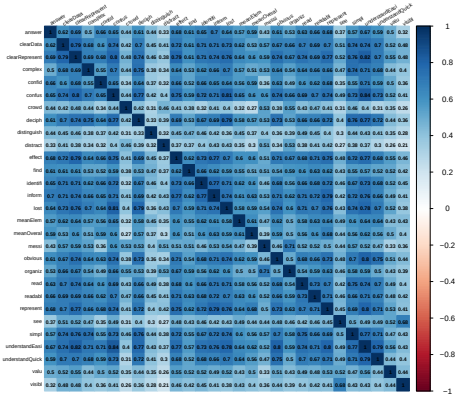

Fig. 19: Correlation matrix from our survey data for **stimulus E** (geneaQuilts) after missing data treatement detailed in Appx. G.

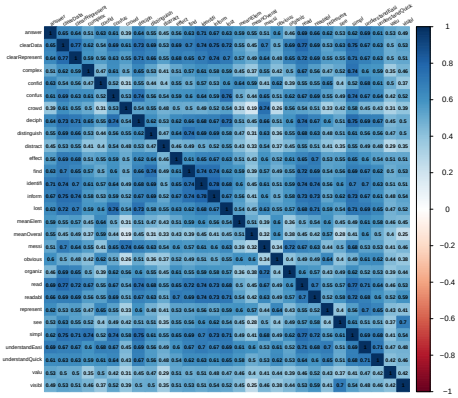

Fig. 20: Correlation matrix from our survey data for **stimulus F** (many lines) after missing data treatement detailed in Appx. G.

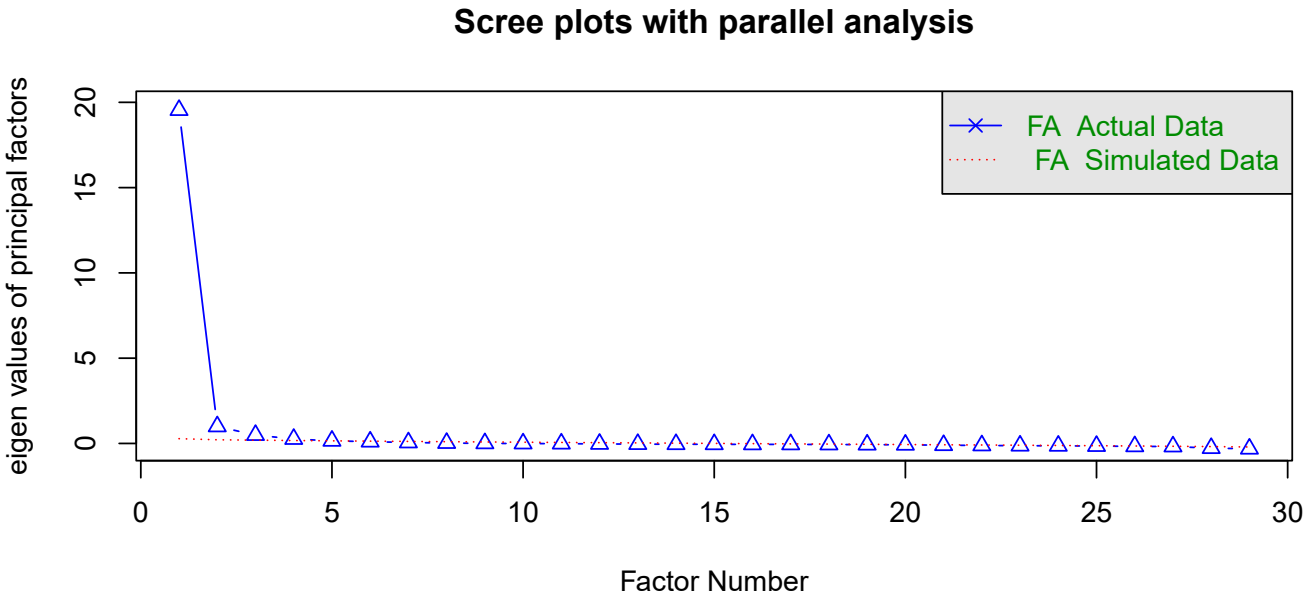

Fig. 21: Scree plot from parallel analysis for our **complete survey data** with actual (blue) and simulated random data (dotted red) lines. Although there is a clear elbow after 1 factor, the blue line and the red dotted line start to overlap at factor 5. This tells us that, although it would be possible to create a unidimensional scale, statistical tests show that 5 factors better explain the data structure. Therefore, before deciding on how many factor we would retain, we explored whether 2-factors, 3-factors, 4-factors and 5-factors solutions could make conceptual sense.

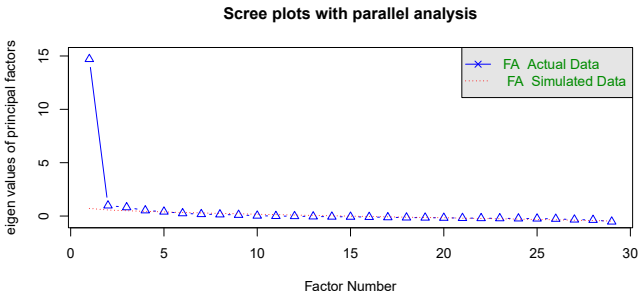

Fig. 22: Scree plot from parallel analysis for **stimulus A**.

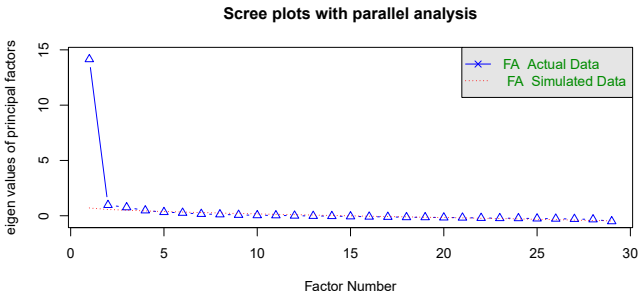

Fig. 23: Scree plot from parallel analysis for **stimulus B**.

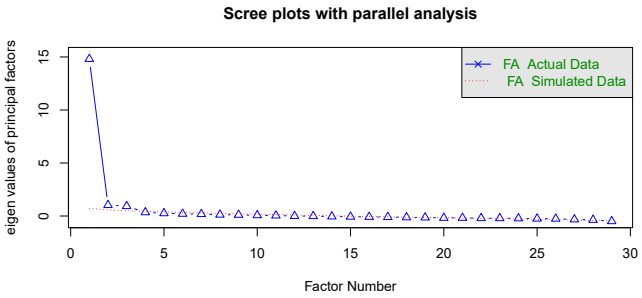

Fig. 24: Scree plot from parallel analysis for **stimulus C**.

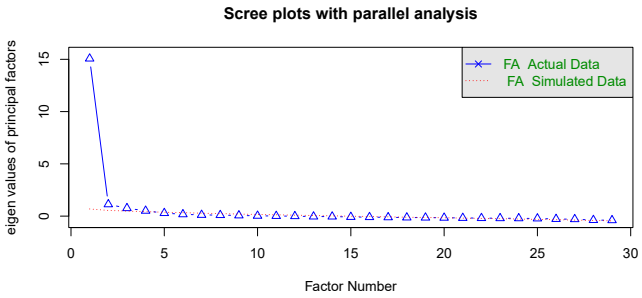

Fig. 25: Scree plot from parallel analysis for **stimulus D**.

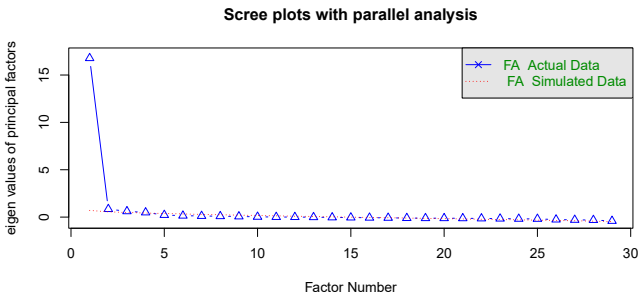

Fig. 26: Scree plot from parallel analysis for **stimulus E**.

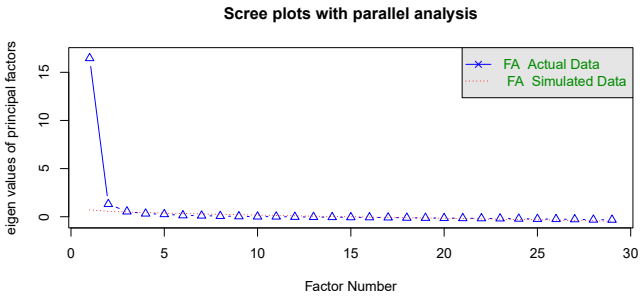

Fig. 27: Scree plot from parallel analysis for **stimulus F**.

K EFA: FACTORS LOADINGS AND NUMBER OF FACTORS TO RETAIN

In this section, we present and discuss factor loading tables generated as part of our Exploratory Factor Analyses in Sect. 5.5.1.

A 1-factor solution (Table 14) would have provided us with the opportunity to develop a simple instrument; furthermore, it seems like a viable candidate according to visual analysis of scree plots (see Appx. J), because all plots show a clear break in the slop after 1 factor. However, there were many limitations in retaining a 1-factor solution for our final instrument:

- A single scale would collapse the amount of information that our instrument can measure on a single average score, which might not help other researchers shed light on their empirical results. As parallel analyses and model fit metrics show, the construct appears to be better explained with 3 to 5 factors.
- As we describe in our OSF Research log, loadings in a 1-factor solution appeared to be very different from one stimulus to another, suggesting that different reasons lead people to find visualization easy or difficult to read depending on the representation they saw. Therefore, deciding on which item should be included in a final, parsimonious scale would require many trade-offs, ultimately lowering the precision and usefulness of the instrument for other researchers.
- As the readability domain lacks formal definition, a detailed instrument would provide more opportunities for further research work on understanding and explaining readability.

Regarding multiple factors solutions, we discarded the 2-factor solution (Table 15) as it did not make much conceptual sense. Similarly, the last factor in the 5-factors structure (Table 18) was vague and weak in explained variance, so we discarded this solution as well.

In the 3-factors structures, we found the grouping of items to be meaningful. The first factor related to the ability of the reader to understand the visualization, with the following 3 first items in terms of loading importance: "It is obvious for me how to read this visualization", "I am confident in my understanding of this visualization", and "I can easily understand the overall meaning of this data visualization". The second factor related to the visual clarity of the layout, with the first 3 following items: "I don't find this visualization crowded", "I don't find this visualization messy", and "I don't find distracting parts in this visualization". Finally, the two items relating to reading data features were set together: "I find data features (for example, a minimum, or an outlier, or a trend) visible in this visualization" and "I can clearly see data features (for example, a minimum, or an outlier, or a trend) in this visualization".

However, some items were distributed across factors in this solution, for which me thought that they might better belong together, in a separate factor. For example: "This visualization effectively shows the data to me" in the first factor and "This visualization shows the data in a clear way for me" in the second factor, or "I can easily read this visualization" in the first factor and "I find this visualization readable" in the second factor. These problems did not appear anymore in the 4-factor structure—in which all items cited in the previous sentence belong to the 4th factor of "DataReading".

Together with model fit indices described in our OSF Research log, we concluded our analysis by choosing the 4-factors solution and conducted a Multi-Group Confirmatory Factor Analysis (MG-CFA) to validate this choice, as described in Appx. L.

Table 14: EFA with **1 factor**: factor loadings for 29 items, and proportion of the total variance in our survey data explained by the factor.

terms	PA1
answer	0.811
clearData	0.915
clearRepresent	0.909
complex	0.852
confid	0.771
confus	0.891
crowd	0.650
deciph	0.877
distinguish	0.717
distract	0.724
effect	0.872
find	0.822
identifi	0.886
inform	0.890
lost	0.879
meanElem	0.795
meanOveral	0.718
messi	0.788
obvious	0.789
organiz	0.815
read	0.907
readabl	0.874
represent	0.842
see	0.707
simpl	0.913
understandEasi	0.888
understandQuick	0.875
valu	0.633
visibl	0.673
Proportion of total variance explained	**0.67**

Table 15: EFA with **2 factors**: factor loadings for 29 items, and proportion of the total variance explained by each factor in our survey data.

items	PA1	PA2
answer	0.756	0.088
clearData	0.402	0.567
clearRepresent	0.536	0.421
complex	0.479	0.419
confid	0.943	-0.145
confus	0.494	0.444
crowd	-0.278	0.994
deciph	0.495	0.429
distinguish	0.038	0.734
distract	-0.065	0.850
effect	0.566	0.350
find	0.312	0.560
identifi	0.585	0.345
inform	0.525	0.411
lost	0.527	0.397
meanElem	0.737	0.090
meanOveral	0.915	-0.174
messi	-0.121	0.980
obvious	1.003	-0.188
organiz	0.162	0.710
read	0.502	0.453
readabl	0.453	0.469
represent	0.926	-0.053
see	0.484	0.258
simpl	0.443	0.521
understandEasi	0.884	0.039
understandQuick	0.793	0.119
valu	0.479	0.183
visibl	0.473	0.233
Proportion of total variance explained	**0.42**	**0.30**

Table 16: EFA with **3 factors**: factor loadings for 29 items, and proportion of the total variance explained by each factor in our survey data.

terms	PA1	PA2	PA3
answer	0.669	0.046	0.145
clearData	0.305	0.520	0.157
clearRepresent	0.458	0.385	0.126
complex	0.507	0.448	-0.059
confid	0.956	-0.135	-0.020
confus	0.533	0.480	-0.078
crowd	-0.211	1.065	-0.151
deciph	0.526	0.460	-0.065
distinguish	-0.120	0.656	0.257
distract	-0.050	0.877	-0.046
effect	0.431	0.279	0.228
find	0.109	0.454	0.340
identifi	0.456	0.279	0.216
inform	0.394	0.344	0.219
lost	0.548	0.421	-0.045
meanElem	0.655	0.050	0.138
meanOveral	0.929	-0.164	-0.020
messi	-0.086	1.026	-0.090
obvious	1.044	-0.169	-0.059
organiz	0.099	0.682	0.099
read	0.441	0.428	0.096
readabl	0.340	0.412	0.186
represent	0.911	-0.056	0.025
see	-0.020	-0.057	0.903
simpl	0.418	0.518	0.032
understandEasi	0.901	0.053	-0.028
understandQuick	0.796	0.127	-0.009
valu	0.195	0.010	0.503
visibl	-0.042	-0.094	0.928
Proportion of total variance	**0.36**	**0.27**	**0.11**

Table 17: EFA with **4 factors**: factor loadings for 29 items, and proportion of the total variance explained by each factor in our survey data.

terms	PA1	PA2	PA4	PA3
answer	0.565	-0.083	0.381	-0.012
clearData	0.255	0.346	0.361	0.039
clearRepresent	0.402	0.246	0.310	0.025
complex	0.503	0.397	0.024	-0.001
confid	0.890	-0.122	0.057	-0.024
confus	0.517	0.407	0.060	-0.026
crowd	-0.135	1.001	-0.113	0.006
deciph	0.476	0.333	0.190	-0.068
distinguish	-0.227	0.353	0.658	0.010
distract	-0.038	0.711	0.158	-0.020
effect	0.360	0.129	0.397	0.061
find	-0.057	0.120	0.819	0.010
identifi	0.322	0.035	0.604	-0.024
inform	0.269	0.097	0.598	-0.015
lost	0.533	0.361	0.057	-0.003
meanElem	0.586	-0.017	0.240	0.040
meanOveral	0.897	-0.097	-0.063	0.019
messi	-0.011	0.974	-0.098	0.049
obvious	1.066	-0.036	-0.242	0.055
organiz	0.133	0.599	0.091	0.105
read	0.335	0.204	0.480	-0.059
readabl	0.242	0.191	0.509	-0.003
represent	0.885	-0.009	-0.026	0.050
see	0.064	0.063	0.071	0.748
simpl	0.385	0.392	0.209	0.004
understandEasi	0.845	0.037	0.071	-0.018
understandQuick	0.770	0.128	0.014	0.023
valu	0.098	-0.157	0.587	0.184
visibl	0.038	0.040	0.039	0.808
Proportion of total variance	**0.32**	**0.19**	**0.19**	**0.06**

Table 18: EFA with **5 factors**: factor loadings for 29 items, and proportion of the total variance explained by each factor in our survey data.

terms	PA1	PA4	PA5	PA2	PA3
answer	0.544	0.401	-0.041	-0.010	-0.021
clearData	0.101	0.330	0.356	0.182	0.051
clearRepresent	0.250	0.286	0.311	0.124	0.034
complex	0.256	-0.024	0.519	0.184	0.014
confid	0.858	0.096	-0.073	-0.008	-0.038
confus	0.317	0.038	0.422	0.231	-0.018
crowd	-0.004	-0.075	-0.032	0.931	-0.003
deciph	0.248	0.150	0.477	0.143	-0.062
distinguish	-0.100	0.693	-0.135	0.359	-0.002
distract	0.061	0.203	-0.035	0.658	-0.034
effect	0.248	0.376	0.212	0.056	0.070
find	-0.009	0.812	-0.028	0.122	0.007
identifi	0.280	0.595	0.074	0.034	-0.026
inform	0.186	0.571	0.177	0.041	-0.012
lost	0.350	0.041	0.373	0.211	0.004
meanElem	0.593	0.277	-0.092	0.068	0.031
meanOveral	0.921	-0.016	-0.165	0.049	0.004
messi	-0.002	-0.060	0.190	0.782	0.048
obvious	0.880	-0.218	0.183	-0.028	0.057
organiz	0.114	0.108	0.146	0.481	0.109
read	0.155	0.440	0.380	0.056	-0.054
readabl	0.152	0.484	0.211	0.107	0.002
represent	0.802	0.006	0.034	0.045	0.046
see	0.021	0.054	0.026	0.024	0.800
simpl	0.128	0.152	0.557	0.154	0.020
understandEasi	0.642	0.063	0.292	-0.020	-0.014
understandQuick	0.528	-0.010	0.401	0.007	0.034
valu	0.080	0.562	-0.001	-0.133	0.195
visibl	0.039	0.038	-0.066	0.037	0.852
Proportion of total variance	**0.25**	**0.18**	**0.14**	**0.13**	**0.06**

Factor loadings analysis — Original file: Agg_EFA_loadings-4_factors.csv

	conceptual meaning >				understanding, meanings		visual clarity (layout quality)		reading the data		reading between the data	
	Factor variable name in R >				*understand*		*layout*		*dataRead*		*dataFeat*	
	Variance proportion explained >				**0.43**		**0.25**		**0.25**		**0.07**	
variable name in R in full model	**statement**	**item**	**Cross-loadings in Agg**	**factor rank in PA1 Agg**	**PA1 Agg**	**factor rank in PA2 Agg**	**PA2 Agg**	**factor rank in PA4 Agg**	**PA4 Agg**	**factor rank in PA3 Agg**	**PA3 Agg**	Order in Agg
understand1	**It is obvious for me how to read this visualization**	**obvious**	**0**	**1**	**1.062337821**							**1**
understand2	**I can easily understand the overall meaning of this data visualization**	**meanOveral**	**0**	**2**	**0.900123582**							**2**
understand3	**I am confident in my understanding of this visualization**	**confid**	**0**	**3**	**0.88850186**							**3**
understand4	**I can easily understand how the data is represented in this visualization**	**represent**	**0**	**4**	**0.882837581**							**4**
understand5	**I can easily understand this visualization**	**understandEasi**	**0**	**5**	**0.845610731**							**5**
understand6	**I can quickly understand this visualization**	**understandQuic**	**0**	**6**	**0.768301622**							**6**
understand7	I can easily understand what the different elements of the visualization mean	meanElem	0	7	0.584213463							7
	I can easily answer some questions about the represented data with this visualization	answer	1	8	0.559170953				0.388855235			8
	I don't feel lost trying to read this visualization	lost	1	9	0.53136155		0.362853201					9
	I don't find this visualization confusing	confus	1	10	0.514550271		0.412462743					10
	I don't find this visualization complex to read	complex	1	11	0.503407051		0.395441902					11
	I don't find the presentation of the data difficult to decipher in this visualization	deciph	1	12	0.472499253		0.331068973					12
	The representation of the data makes the information clear to me in this visualization	clearRepresent	0	13	0.401080111							13
layout1	**I don't find this visualization crowded**	**crowd**	**0**			**1**	**1.002355239**					**14**
layout2	**I don't find this visualization messy**	**messi**	**0**			**2**	**0.97629751**					**15**
layout3	**I don't find distracting parts in this visualization**	**distract**	**0**			**3**	**0.716152421**					**16**
layout4	I find this visualization well organized	organiz	0			4	0.601772217					17
	I find this visualization simple to read	simpl	1		0.380395624	5	0.393556178					18
dataFeat1	**I find data features (for example, a minimum, or an outlier, or a trend) visible in this visualization**	**visibl**	**0**							**1**	**0.801035197**	**19**
dataFeat2	**I can clearly see data features (for example, a minimum, or an outlier, or a trend) in this visualization**	**see**	**0**							**2**	**0.745991576**	**20**
dataRead1	**I can easily find specific elements in this visualization**	**find**	**0**					**1**	**0.813425997**			**21**
	I can easily distinguish individual elements of the represented data (for example individual lines, or dots, or areas, or colors...)	distinguish	1				0.355062313	2	0.655425343			22
dataRead2	I can easily identify relevant information in this visualization	identifi	0					3	0.603689212			23
dataRead3	I can read data values from this visualization	valu	0					4	0.600135661			24
dataRead4	I can easily retrieve information from this visualization	inform	0					5	0.599757226			25
dataRead5	I find this visualization readable	readabl	0					6	0.505646165			26
	I can easily read this visualization	read	1		0.333193092			7	0.484235378			27
	This visualization effectively shows the data to me	effect	1		0.35791681			8	0.399574603			28
	This visualization shows the data in a clear way for me	clearData	1				0.347569591	9	0.36121256			29

Fig. 28: Table of factors loadings generated for the 4-factors solution found with a Principal Axis method. Fully colored lines indicate items that we screened as candidates for further tests of model fit and reliability, because they met the following selection criteria: items should not have loadings > 0.3 in more than 1 factor (no cross-loadings), and items should have a loading value > 0.5 for their associated factor (strong loading value). This table was generated following 3 steps. We provide all data and code in the supplemental material for steps 1 and 2, and a detailed explanation for all 3 steps in this description. **Step 1: Generate raw factor loadings.** We used our `EFA-+-all.Rmd` notebook in `R` to conduct EFA on the collected data. We generate a raw factor loading table for a 4-factors structure with the Principal Axis (PA) method `PA1 Stimuli 4 factors - Agg.csv`. **Step 2: Factor loadings pre-processing.** We used this table as an input for our `EFA factor loading analysis.ipynb` notebook in `Python` for pre-analysis processing. The pre-processing consisted in 3 interventions: a) filtering poor loadings values out, b) ranking the items for easier ordering of the table, and c) managing cross-loadings. In a) we replaced all loadings < 0.3 with `NA` values. In b) we created the 4 "factor ranks" columns, so that, for each factor, all items item belonging to the factor were ranked by loading importance (1 being the highest). When an item exhibited loadings > 0.3 for multiple factors, we assigned it to the factor where its loading was the highest (e. g., in the table above, the item "lost" had loadings >0.3 in both factors PA1 and PA2, but its factor loading value was higher for PA1; therefore it was assigned to PA1 for ranking). Additionally, we created the "Order in Agg" to facilitate the sorting of the complete spreadsheet, by incrementing all ranks factor after factor. In c) we counted, for each item, how many times it cross-loaded (number of factors after the first where the item had a loading > 0.3) and saved this value in the "Cross-loadings in Agg" column. **Step 3: Table visual formatting and annotations (manual).** To facilitate analysis and discussion, we created a spreadsheet with the following layout rules: a) yellow background where cross-loadings counts were not null, b) bold face on items with factor loading values > 0.7, and c) red color on factor loading values < 0.5. We imported the statements corresponding to item codes to facilitate our conceptual analysis and reported the variance proportion explained from the EFA `R` output. We then annotated the table with our understanding of chich concepts the 4 factors represented, based on the associated items. We attributed one color to each factor to visually bring the structure up. Finally, we fully colored lines for items that met our selection criteria for the next step of item reduction. *Note: steps 1 and 2 were conducted for 1-factor, 2-factors, 3-factors and 4-factors solutions. For 3-factors and 4-factors solutions, we also conducted EFA with the Maximum-Likelihood and observed that the factor loadings were not significantly different. We performed step 3) only for 3-factors adn and 4-factors solutions.*

L Multi-Group Confirmatory Factor Analysis with a 4-factors model

Confirmatory Factor Analysis (CFA) is used to examine goodness of fit of a model to a dataset by comparing it to a baseline (null) model. Multi-Group CFA (MG-CFA) is a variation of CFA with an added grouping variable (in our case, the 6 independent groups of participant who rated a specific visualization stimulus). MG-CFA allows to assess to what extent the factor structure (patterns of factor loadings and factor covariances) is consistent across different groups of respondents. The output of CFA is examined through the model fit metrics it produces.

We performed MG-CFA using the `lavaan` package in R with the following full 4-factor structure extracted from Fig. 28:

```
full_Readability_factors <- list(
  understand = c('obvious', 'meanOveral', 'confid',
                 'represent', 'understandEasi',
                 'understandQuick', 'meanElem'),
  layout = c('crowd', 'messi', 'distract', 'organiz'),
  dataRead = c('find', 'identifi', 'valu', 'inform',
               'readabl'),
  dataFeat = c('visibl', 'see')
)
```

We report a partial view of resulting fit metrics in Table 19, and we share the R notebook as well as the complete output in our supplemental material. Notably, we obtained the following metrics:

- **Measures of fit indices** (values closer to 1 are better): the Tucker–Lewis Index (TLI) was .94 and the Comparative Fit Index (CFI) was .95;
- **Measures of covariance discrepancies** between observed and the model-implied data (values closer to 0 are better): The Standardized Root Mean square Residual (SRMR) was .046 and the Root Mean Square Error of Approximation (RMSEA) was .067.

In their reference work on cutoff criteria for fit indices, Hu and Bentler [50] recommend a combination of two criteria to retain a model: fit indices such as TLI or CFI should be higher than .95, and SRMR should be lower than 0.9. They add that a combinational cutoff criterion of RMSEA at .06 and SRMR at .09 is possible but less desirable because it tends to reject models that are, in fact, good fits.

With those results, we were satisfied that our 4-factors model was a good fit to explain our survey's data. However, CFA conducted on the same dataset from which the factors were extracted is expected to produce good results, and further validation is required by conducting CFA on a set of observations from an independent group [12]. We conducted such analysis in the validation phase of our work in Sect. 6.3.

M Item-subscale reliability

We provide in Table 20 reliability and related statistics for each individual item from the 4-factors structure we identified in Appx. K and tested in Appx. L. These values inform us about the importance of each item regarding in ensuring reliability in measures of the factor it relates to. We use these values to build model_2 in the following Appx. N.

N Reducing items in subscales

In this section we provide details regarding the final step of scale development: the selection of items that will compose our final subscales.

Reference authors emphasize that item reduction is the heart of scale development [30]. It entails the use of tools from Classical Test Theory (CTT) and Item Response Theory (IRT). While IRT has a more refined approach in assessing items characteristics that might affect their measuring performance, IRT analysis for less than 10 items might be unreliable [104]. We did run an IRT analysis, with the goal to provide additional information in case other decision criteria appeared inconclusive, but we only used it to confirm our final selection of subscales' items.

In CTT, researchers use reliability indicators such as Cronbach's alpha to select items that will minimize the final scale's measurement error [30]. Additionally, researchers can consider factor loadings [12] for selecting items that capture a good amount of information about the factor's underlying construct.

Table 19: Fit metrics from Multi-Group Confirmatory Factor Analysis at the scale development stage. In this analysis we fit the full 4-factors model extracted from our EFA to our survey data, using the stimulus as grouping variable. The resulting fit metrics allow us to estimate how appropriate the 4-factor structure is for explaining not only the entire aggregate set of answers, but also subsets of data in our independent groups, as each participant had been randomly assigned to rate 1 of our 6 stimuli visualizations.

Fit metric	Value from CFA
chisq	1802
df	774
pvalue	0
baseline.chisq	21465
baseline.df	918
baseline.pvalue	0
cfi	0.950
tli	0.941
cfi.robust	0.950
tli.robust	0.940
rmsea	0.067
rmsea.ci.lower	0.063
rmsea.ci.upper	0.071
rmsea.ci.level	0.9
rmsea.pvalue	4.2886e-12
srmr	0.0456
srmr_bentler	0.0456
srmr_bentler_nomean	0.0479
crmr	0.0479
crmr_nomean	0.0506
srmr_mplus	0.0456
srmr_mplus_nomean	0.0479

N.1 Creating 3 combinations of items

To explore possible combinations of items, we calculated items ranks based on factor loadings and on reliability indicators for each subscale. We produced 3 candidate combinations of items: a reliability optimizer, a factor loadings optimizer, and a mixed approach based on average ranks. For reliability optimization (model 1), we calculated the effect of dropping items on reliability indicators for each scale using the `psych` package in R. The lower the alpha drops if an item were removed, the more important the item is to ensure reliability of the final scale. We thus ranked the items based on how much the scale' alpha would lowered if they were dropped. For loadings optimization (model 2), we retrieved the factors loadings from our EFA. We then ranked each item accordingly to their loading values, higher being better. For each item, we also calculate an average of the two ranks and use it as a basis for a mixed approach (model 3), which might offer a useful trade-off between reliability and factor representation.

For each approach, we selected the 3 higher ranking items in UNDERSTAND, LAYOUT, and DATAREAD to build a model. DATAFEAT has only two items which remain the same across all models. We then put the 3 models to test. We provide the ranks and the selected items in Table 21.

N.2 Comparing results: reliability and model fit

For each of the produced models, we used the `psych` package to calculate Cronbach's alpha and McDonald's omega. We provide the code and data sources for this analysis in our supplemental material.

Table 22 to Table 24 show Cronbach's alpha coefficients calculated with `R's psych` package for each option. We used the `lavaan` package for conducting Multi-Group Confirmatory Factor Analysis (MG-CFA) for each model, and compared model fit values. We provide code and complete data from this analysis in our supplemental material.

Table 20: Reliability and related statistics for each individual item. It provides reliability metrics for each factor, if an item were removed. In such a table, the more an alpha, G6, r (items correlation), or S/N value drops, the more important an item is to a given factor's reliability.

Items	Factor	raw_alpha	std.alpha	alpha se	G6(smc)	var.r	med.r	average_r	S/N
obvious	understand	0.941	0.943	0.002	0.935	0.003	0.728	0.733	16.459
meanOveral	understand	0.947	0.948	0.002	0.941	0.002	0.740	0.753	18.314
confid	understand	0.944	0.945	0.002	0.938	0.003	0.731	0.743	17.332
represent	understand	0.939	0.941	0.002	0.933	0.003	0.719	0.725	15.817
understandEasi	understand	0.937	0.939	0.002	0.930	0.002	0.719	0.719	15.384
understandQuick	understand	0.941	0.942	0.002	0.933	0.002	0.731	0.729	16.101
meanElem	understand	0.946	0.947	0.002	0.940	0.002	0.740	0.750	18.019
crowd	layout	0.887	0.887	0.005	0.846	0.004	0.723	0.724	7.869
messi	layout	0.855	0.856	0.006	0.798	0.000	0.665	0.664	5.933
distract	layout	0.891	0.892	0.005	0.856	0.005	0.762	0.733	8.223
organiz	layout	0.885	0.885	0.005	0.840	0.002	0.723	0.720	7.708
find	dataRead	0.900	0.899	0.004	0.883	0.016	0.683	0.690	8.899
identifi	dataRead	0.895	0.894	0.004	0.874	0.011	0.675	0.678	8.429
valu	dataRead	0.935	0.936	0.003	0.917	0.001	0.778	0.784	14.505
inform	dataRead	0.893	0.893	0.004	0.871	0.010	0.675	0.675	8.310
readabl	dataRead	0.900	0.899	0.004	0.883	0.013	0.685	0.691	8.945
visibl	dataFeat	0.765	0.784	NA	0.614	0.000	0.784	0.784	3.627
see	dataFeat	0.804	0.784	NA	0.614	0.000	0.784	0.784	3.627

Table 21: Ranking of candidate items for each subscale by factor loading and reliability indices to build 3 possible models based on: rank in factor loadings, rank in effect on reliability if dropped, and an average of the two ranks.

Model 3	**Model 2**	**Model 1**	**Items**	**Factor loadings**	**alpha with all items**	**Reliability of subscale if item was dropped**	
Mean ranking	Effect on reliability if dropped rank	Factor loadings rank	factor_name.item_name	From EFA	raw_alpha of full factor	raw_alpha	std.alpha
2.5 ✓	4	**1** ✓	understand.obvious	1.062	0.950	0.941	0.94
3 ✓	**2** ✓	4	understand.represent	0.883	0.950	0.939	0.94
3 ✓	**1** ✓	5	understand.understandEasi	0.846	0.950	0.937	0.94
4	6	**2** ✓	understand.meanOveral	0.900	0.950	0.947	0.95
4	5	**3** ✓	understand.confid	0.889	0.950	0.944	0.95
4.5	**3** ✓	6	understand.understandQuick	0.768	0.950	0.941	0.94
7	7	7	understand.meanElem	0.584	0.950	0.946	0.95
1.5 ✓	**1** ✓	**2** ✓	layout.messi	0.976	0.907	0.855	0.86
2 ✓	**3** ✓	**1** ✓	layout.crowd	1.002	0.907	0.887	0.89
3 ✓	**2** ✓	4	layout.organiz	0.602	0.907	0.885	0.89
3.5	4	**3** ✓	layout.distract	0.716	0.907	0.891	0.89
2 ✓	**3** ✓	**1** ✓	dataRead.find	0.813	0.923	0.900	0.90
2 ✓	**2** ✓	**2** ✓	dataRead.identifi	0.604	0.923	0.895	0.89
2.5 ✓	**1** ✓	4	dataRead.inform	0.600	0.923	0.893	0.89
3.5	4	**3** ✓	dataRead.valu	0.600	0.923	0.935	0.94
4	3	5	dataRead.readabl	0.506	0.923	0.900	0.90
1 ✓	**1** ✓	**1** ✓	dataFeat.visibl	0.801	0.879	0.765	0.78
1.5 ✓	**1** ✓	**2** ✓	dataFeat.see	0.746	0.879	0.804	0.78

Table 22: Cronbach's alpha and McDonald's omega reliability coefficients for Model 1 (reliability-based) in Appx. N

	Subscale	A	B	C	D	E	F	Full survey
omega tot	dataFeat	0.870	0.826	0.837	0.858	0.807	0.822	0.879
	dataRead	0.758	0.741	0.783	0.767	0.804	0.820	0.855
	layout	0.768	0.845	0.838	0.752	0.753	0.822	0.886
	understand	0.811	0.755	0.803	0.814	0.820	0.820	0.878
raw alpha	dataFeat	0.868	0.824	0.835	0.859	0.815	0.822	0.879
	dataRead	0.736	0.739	0.773	0.741	0.802	0.809	0.850
	layout	0.760	0.844	0.834	0.748	0.751	0.801	0.885
	understand	0.809	0.745	0.796	0.812	0.821	0.818	0.874
std alpha	dataFeat	0.868	0.825	0.835	0.859	0.816	0.823	0.879
	dataRead	0.757	0.740	0.776	0.749	0.802	0.808	0.849
	layout	0.765	0.845	0.837	0.747	0.751	0.810	0.885
	understand	0.811	0.747	0.801	0.813	0.823	0.820	0.878

Table 23: Cronbach's alpha and McDonald's omega reliability coefficients for Model 2 (loadings-based) in Appx. N

	Subscale	A	B	C	D	E	F	Full survey
omega tot	dataFeat	0.872	0.823	0.839	0.857	0.811	0.822	0.877
	dataRead	0.828	0.799	0.819	0.870	0.870	0.903	0.920
	layout	0.792	0.850	0.847	0.791	0.821	0.867	0.894
	understand	0.858	0.875	0.874	0.852	0.910	0.868	0.931
raw alpha	dataFeat	0.868	0.824	0.835	0.859	0.815	0.822	0.879
	dataRead	0.820	0.791	0.811	0.870	0.868	0.904	0.920
	layout	0.773	0.841	0.827	0.789	0.815	0.854	0.891
	understand	0.855	0.858	0.869	0.848	0.909	0.865	0.927
std alpha	dataFeat	0.868	0.825	0.835	0.859	0.816	0.823	0.879
	dataRead	0.826	0.794	0.817	0.870	0.868	0.904	0.920
	layout	0.774	0.850	0.837	0.789	0.814	0.860	0.892
	understand	0.857	0.871	0.874	0.849	0.910	0.868	0.930

Table 24: Cronbach's alpha and McDonald's omega reliability coefficients for Model 3 (average ranks from model 1 and 2) in Appx. N

	Subscale	A	B	C	D	E	F	Full survey
omega tot	dataFeat	0.872	0.822	0.838	0.858	0.815	0.822	0.878
	dataRead	0.828	0.799	0.819	0.869	0.871	0.903	0.920
	layout	0.792	0.850	0.847	0.791	0.821	0.868	0.895
	understand	0.869	0.859	0.838	0.839	0.914	0.848	0.923
raw alpha	dataFeat	0.868	0.824	0.835	0.859	0.815	0.822	0.879
	dataRead	0.820	0.791	0.811	0.870	0.868	0.904	0.920
	layout	0.773	0.841	0.827	0.789	0.815	0.854	0.891
	understand	0.867	0.852	0.826	0.837	0.914	0.847	0.921
std alpha	dataFeat	0.868	0.825	0.835	0.859	0.816	0.823	0.879
	dataRead	0.826	0.794	0.817	0.870	0.868	0.904	0.920
	layout	0.774	0.850	0.837	0.789	0.814	0.860	0.892
	understand	0.868	0.853	0.830	0.838	0.914	0.848	0.922

Table 25: Fit metrics from Multi-Group Confirmatory Factor Analysis on our 3 candidate models, as described in Sect. N.1.

Fit indices	Full model (all items)	Model_1	Model_2	Model_3	Model_final
chisq	1802	502	540	537	478
cfi	0.950	0.970	0.974	0.974	0.978
tli	0.941	0.956	0.963	0.962	0.968
srmr	0.046	0.040	0.036	0.036	0.035
rmsea	0.067	0.064	0.068	0.068	0.061

Results showed that all models performed similarly, although the 3d model from average rankings performed slightly better(see Model 1, Model 2 and Model 3 in Table 25). We provide the comparison of all 3 aforementioned models as well as the "full" version (containing all items selected from), the final, adjusted model in Table 25.

N.3 Final selection of items

We used model 3 as our final selection, with one modification: in the LAYOUT subscale, we exchanged the item "I find this visualization well organized" with "I don't find distracting parts in this visualization" in order to ensure better phrasing consistency with other items of the subscale. We obtained the finat set of items described in Table 26.

Table 26: Our final selection of items forming the PREVis instrument across 4 subscales: ◆ **UNDERSTAND**, ◆ **LAYOUT**, ◆ **DATAFEAT**, and ◆ **DATAREAD**.

Item code	Item statement
◆ obvious	It is obvious for me how to read this visualization
◆ represent	I can easily understand how the data is represented in this visualization
◆ understandEasi	I can easily understand this visualization
◆ messi	I don't find this visualization messy
◆ crowd	I don't find this visualization crowded
◆ distract	I don't find parts of the visualization distracting
◆ visibl	I find data features (for example, a minimum, or an outlier, or a trend) visible in this visualization
◆ see	I can clearly see data features (for example, a minimum, or an outlier, or a trend) in this visualization
◆ inform	I can easily retrieve information from this visualization
◆ identifi	I can easily identify relevant information in this visualization
◆ find	I can easily find specific elements in this visualization

We submitted our final selection to the same MG-CFA and obtained improved model fit as shown in Table 25. The reliability measures for this final subscales composition are displayed in Table 27.

O RATINGS PLOTS IN EXPLORATORY SURVEY

In this section we present average ratings of stimuli with 95% CI from our exploratory survey:

- aggregate ratings from each of our final ◆◆◆◆ subscales in Fig. 29;
- for each subscale and each stimulus, the subscale's average rating with the subscale's aggregate score; and
- individual item averages across all stimuli for all items in the original 4-factors structure identified in Appx. L—including items that we did not retain for the final PREVis instrument.

Table 27: Cronbach's alpha and McDonald's omega reliability coefficients for our final subscales ◆◆◆◆.

	Subscale	A	B	C	D	E	F	Full survey
omega tot	dataFeat	0.872	0.826	0.837	0.858	0.811	0.822	0.879
	dataRead	0.829	0.799	0.819	0.870	0.871	0.904	0.920
	layout	0.768	0.844	0.838	0.751	0.753	0.823	0.886
	understand	0.869	0.859	0.839	0.839	0.914	0.849	0.922
raw alpha	dataFeat	0.868	0.824	0.835	0.859	0.815	0.822	0.879
	dataRead	0.820	0.791	0.811	0.870	0.868	0.904	0.920
	layout	0.760	0.844	0.834	0.748	0.751	0.801	0.885
	understand	0.867	0.852	0.826	0.837	0.914	0.847	0.921
std alpha	dataFeat	0.868	0.825	0.835	0.859	0.816	0.823	0.879
	dataRead	0.826	0.794	0.817	0.870	0.868	0.904	0.920
	layout	0.765	0.845	0.837	0.747	0.751	0.810	0.885
	understand	0.868	0.853	0.830	0.838	0.914	0.848	0.922

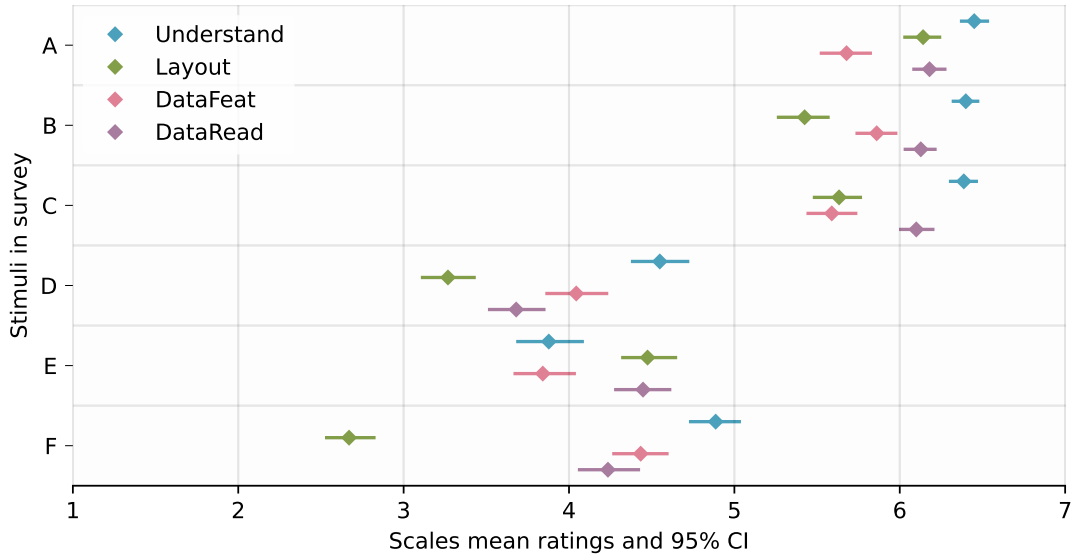

Fig. 29: Average ratings (from 1 = "Strongly disagree" to 7 = "Strongly agree") using the four PREVis subscales ◆◆◆◆ on 6 visualizations of different readability: (A / B / C) > (D / E / F). For a given subscale, ratings and 95% CI do not overlap and fit the readability ranking.

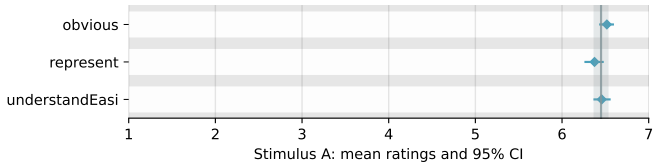

Fig. 30: Comparison of average ratings from ◆ **UNDERSTAND** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **A**.

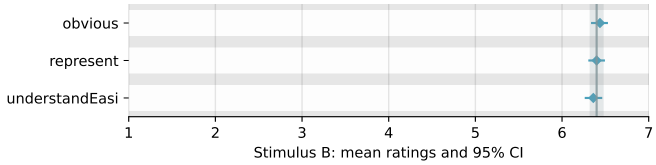

Fig. 31: Comparison of average ratings from ◆ **UNDERSTAND** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **B**.

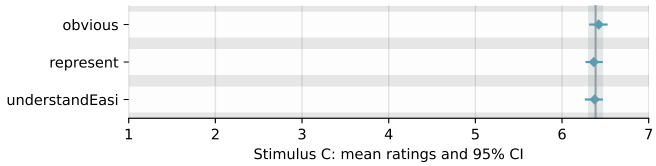

Fig. 32: Comparison of average ratings from ◆ **UNDERSTAND** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **C**.

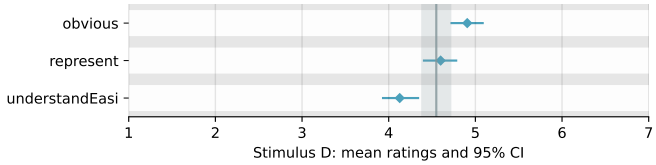

Fig. 33: Comparison of average ratings from ◆ **UNDERSTAND** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **D**.

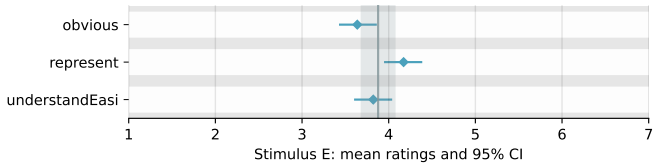

Fig. 34: Comparison of average ratings from ◆ **UNDERSTAND** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **E**.

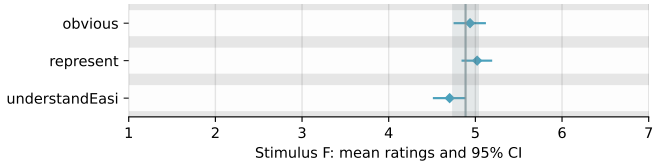

Fig. 35: Comparison of average ratings from ◆ **UNDERSTAND** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **F**.

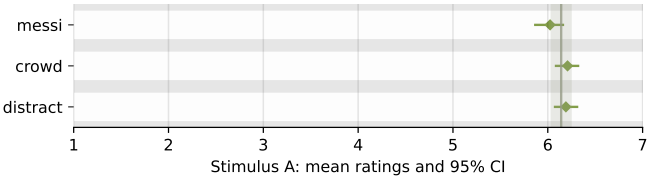

Fig. 36: Comparison of average ratings from ◆ **LAYOUT** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **A**.

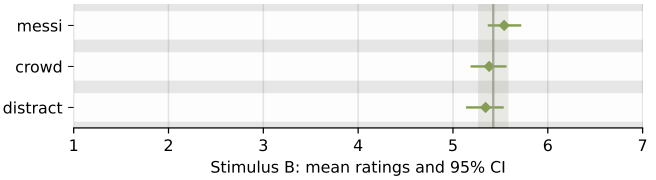

Fig. 37: Comparison of average ratings from ◆ **LAYOUT** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **B**.

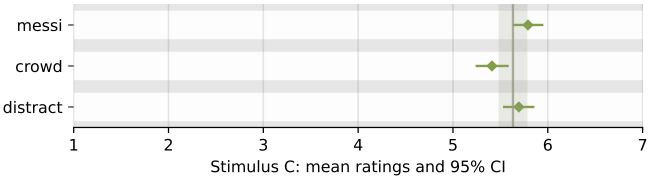

Fig. 38: Comparison of average ratings from ◆ **LAYOUT** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **C**.

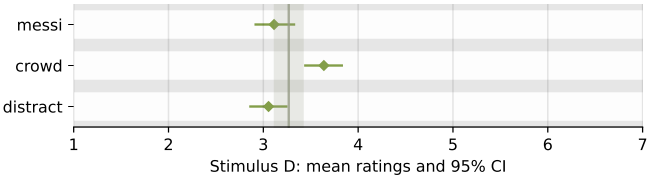

Fig. 39: Comparison of average ratings from ◆ **LAYOUT** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **D**.

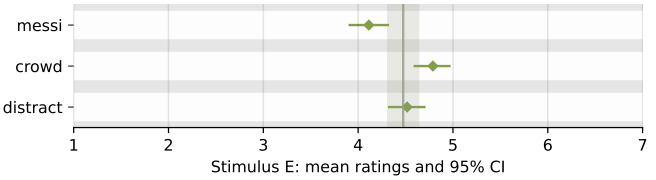

Fig. 40: Comparison of average ratings from ◆ **LAYOUT** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **E**.

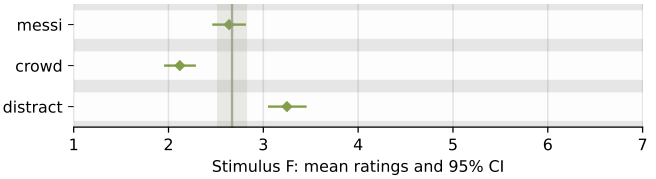

Fig. 41: Comparison of average ratings from ◆ **LAYOUT** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **F**.

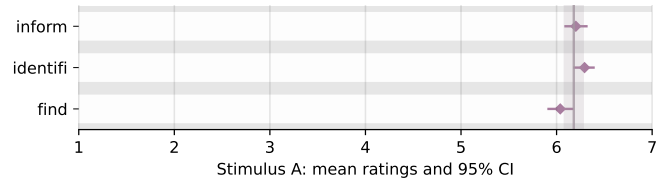

Fig. 42: Comparison of average ratings from ◆ **DataRead** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **A**.

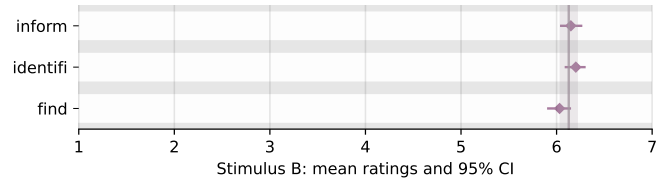

Fig. 43: Comparison of average ratings from ◆ **DataRead** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **B**.

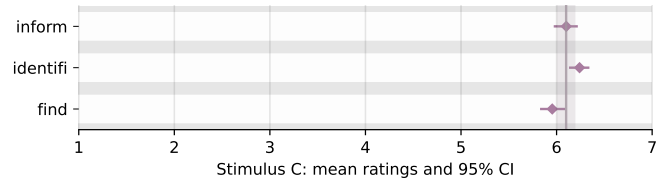

Fig. 44: Comparison of average ratings from ◆ **DataRead** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **C**.

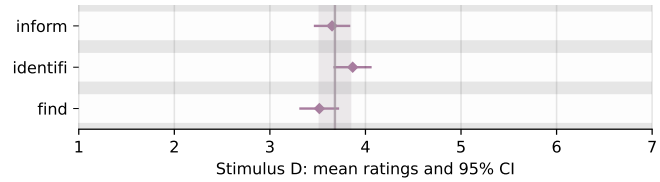

Fig. 45: Comparison of average ratings from ◆ **DataRead** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **D**.

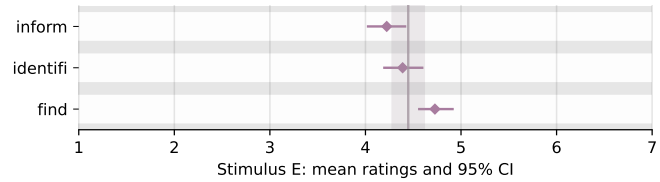

Fig. 46: Comparison of average ratings from ◆ **DataRead** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **E**.

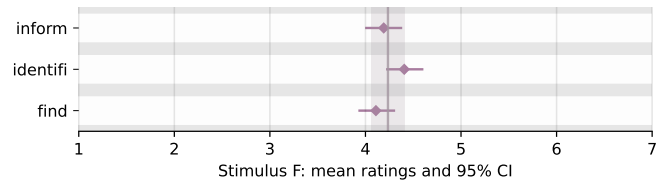

Fig. 47: Comparison of average ratings from ◆ **DataRead** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **F**.

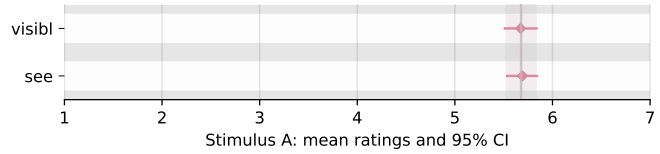

Fig. 48: Comparison of average ratings from ◆ **DataFeat** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **A**.

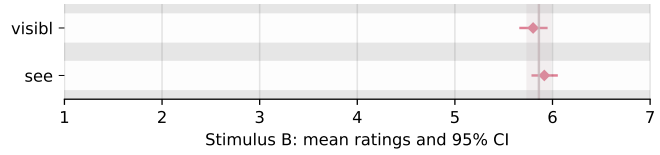

Fig. 49: Comparison of average ratings from ◆ **DataFeat** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **B**.

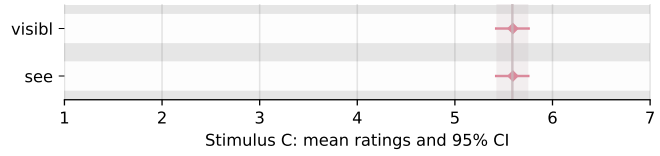

Fig. 50: Comparison of average ratings from ◆ **DataFeat** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **C**.

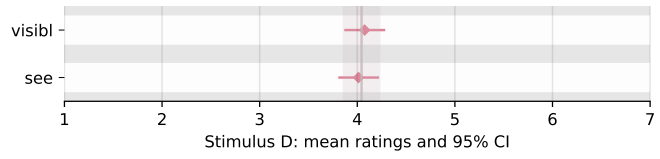

Fig. 51: Comparison of average ratings from ◆ **DataFeat** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **D**.

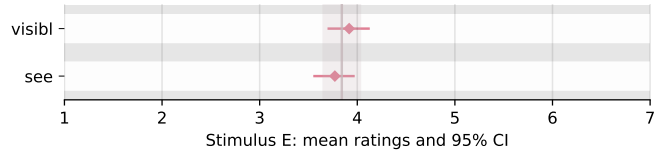

Fig. 52: Comparison of average ratings from ◆ **DataFeat** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **E**.

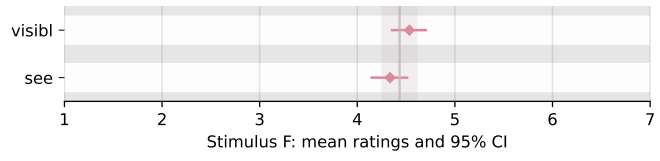

Fig. 53: Comparison of average ratings from ◆ **DataFeat** items, compared to the subscale's average values and 95% CI (vertical line and rectangle) for stimulus **F**.

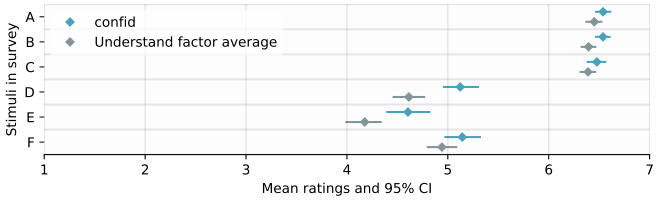

Fig. 54: Comparison of ratings from the ***confid*** item and average ratings from all items in the *Understand* factor, across 6 stimuli in the exploratory survey the.

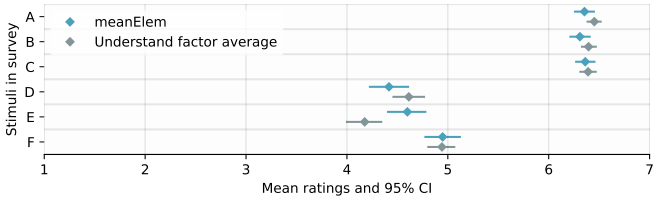

Fig. 55: Comparison of ratings from the ***meanElem*** item and average ratings from all items in the *Understand* factor, across 6 stimuli in the exploratory survey the.

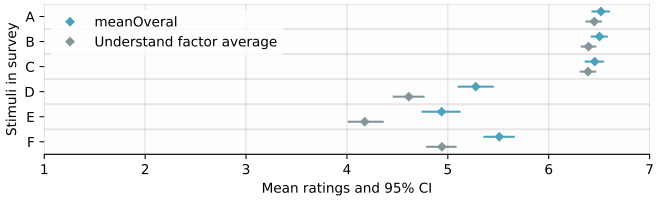

Fig. 56: Comparison of ratings from the ***meanOveral*** item and average ratings from all items in the *Understand* factor, across 6 stimuli in the exploratory survey the.

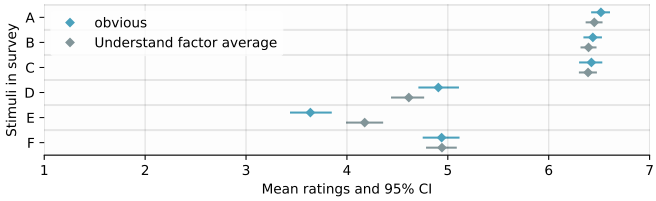

Fig. 57: Comparison of ratings from the ***obvious*** item and average ratings from all items in the *Understand* factor, across 6 stimuli in the exploratory survey the.

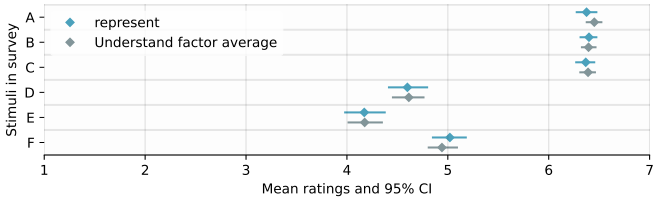

Fig. 58: Comparison of ratings from the ***represent*** item and average ratings from all items in the *Understand* factor, across 6 stimuli in the exploratory survey the.

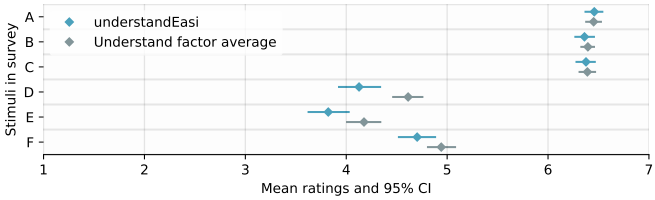

Fig. 59: Comparison of ratings from the ***understandEasi*** item and average ratings from all items in the *Understand* factor, across 6 stimuli in the exploratory survey the.

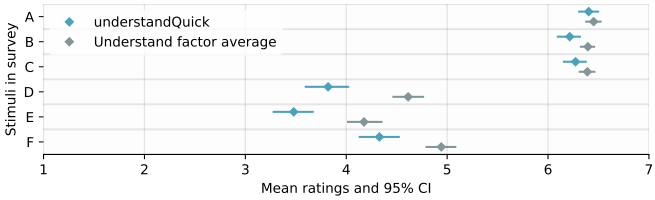

Fig. 60: Comparison of ratings from the ***understandQuick*** item and average ratings from all items in the *Understand* factor, across 6 stimuli in the exploratory survey.

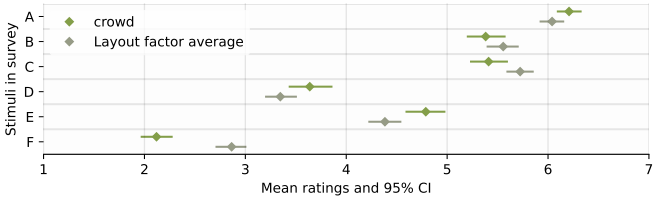

Fig. 61: Comparison of ratings from the ***crowd*** item and average ratings from all items in the *Layout* factor, across 6 stimuli in the exploratory survey the.

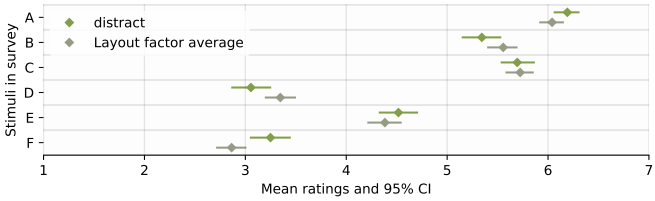

Fig. 62: Comparison of ratings from the ***distract*** item and average ratings from all items in the *Layout* factor, across 6 stimuli in the exploratory survey the.

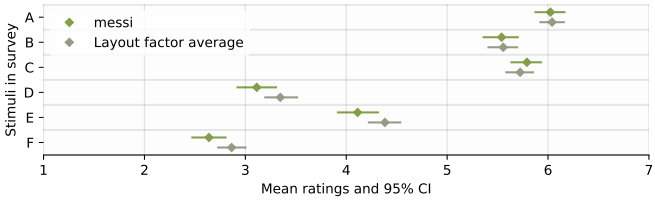

Fig. 63: Comparison of ratings from the *messi* item and average ratings from all items in the *Layout* factor, across 6 stimuli in the exploratory survey the.

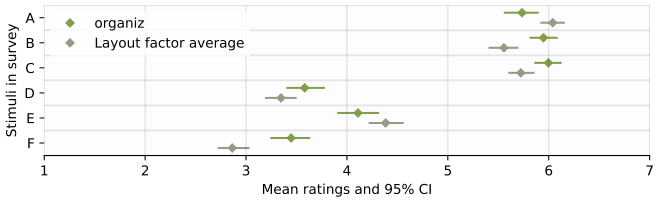

Fig. 64: Comparison of ratings from the ***organiz*** item and average ratings from all items in the *Layout* factor, across 6 stimuli in the exploratory survey the.

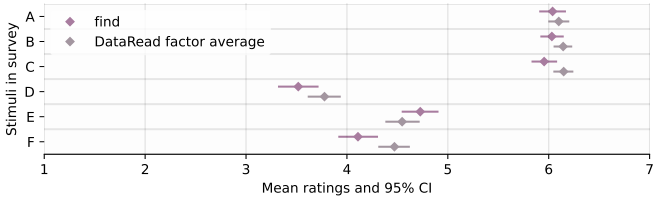

Fig. 65: Comparison of ratings from the ***find*** item and average ratings from all items in the *DataRead* factor, across 6 stimuli in the exploratory survey the.

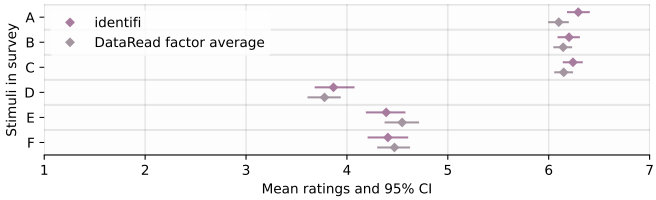

Fig. 66: Comparison of ratings from the ***identifi*** item and average ratings from all items in the *DataRead* factor, across 6 stimuli in the exploratory survey the.

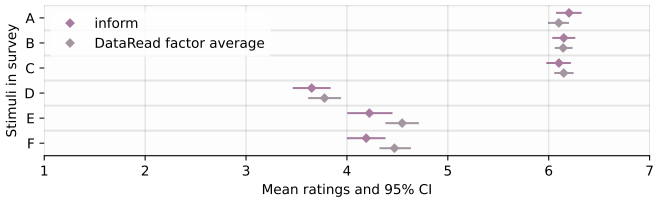

Fig. 67: Comparison of ratings from the ***inform*** item and average ratings from all items in the *DataRead* factor, across 6 stimuli in the exploratory survey the.

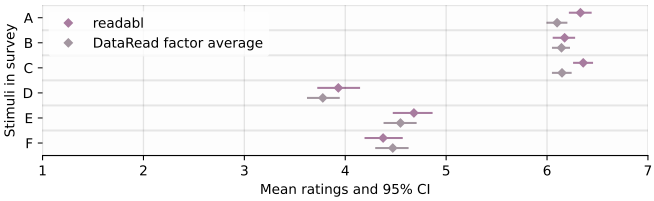

Fig. 68: Comparison of ratings from the ***readabl*** item and average ratings from all items in the *DataRead* factor, across 6 stimuli in the exploratory survey the.

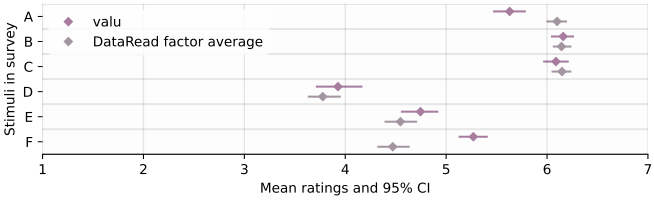

Fig. 69: Comparison of ratings from the ***valu*** item and average ratings from all items in the *DataRead* factor, across 6 stimuli in the exploratory survey the.

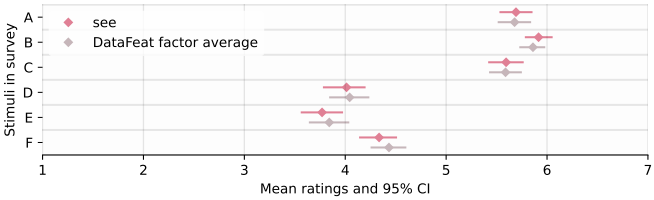

Fig. 70: Comparison of ratings from the ***see*** item and average ratings from all items in the *DataFeat* factor, across 6 stimuli in the exploratory survey the.

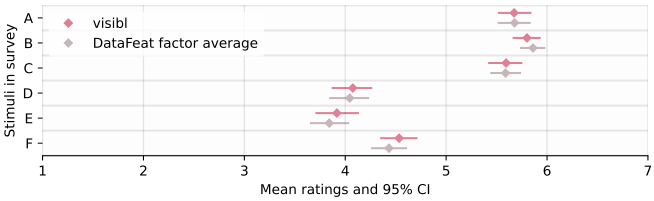

Fig. 71: Comparison of ratings from the ***visibl*** item and average ratings from all items in the *DataFeat* factor, across 6 stimuli in the exploratory survey the.

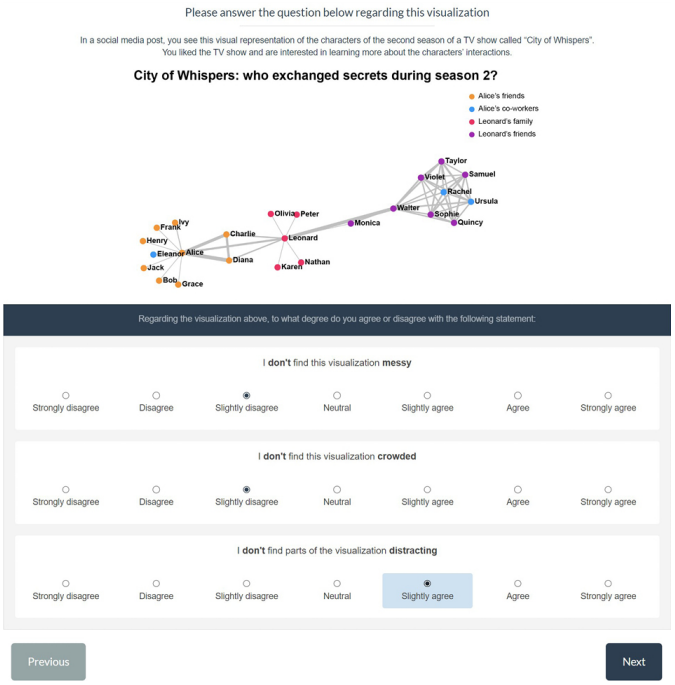

Fig. 72: Example screenshot from our validation survey in action: rating stimulus node-link visualization Ⓑ with ◆ UNDERSTAND items.

P VALIDATION SURVEY DESIGN

As we stated in Sect. 6.1, our validation study design was mainly based on the exploratory study described in Sect. 5.1, but we made some adjustments. Here, we detail the reason for these changes.

All participants rated all stimuli. This choice was motivated by the fact that future researchers might be interested in running studies where participants rate multiple stimuli. As we had an independent-groups design for our exploratory survey, we decided to run a within-participants study.

Rating items for each subscale were presented together on a single screen. We expected our study to exceed the 10 minutes target, after which crowd-sourced work can induce fatigue in participants; therefore, we were seeking ways to facilitate participant's task as much as possible. Showing groups of ratings instead of a single item per screen was a solution already successfully implemented in our team's previous work [46] and showed good time performance. Fig. 72 shows an example screenshot from our survey design. However, with this style of questions in the LimeSurvey platform we used, we lost the possibility to provide an "I don't know / Not applicable" answer option with an optional open text field (as we had done in the exploratory survey). Instead, we added a separate, optional comment question for participants who wanted to qualify their answers about each visualization they rated.

We used 3 stimuli of a single type (node-link). As explained in Sect. 6.1, in order to produce our Multi-Trait Multi-Method matrix (Fig. 4), we decided to compare PREVis ◆◆◆◆ scores with graph aesthetics metrics from Gove's work [43] to assess convergent validity of our final instrument. As a result, we only used node-link diagrams. We also wanted to verify that average ◆◆◆◆ scores with 95% CI allowed us to distinguish between different levels of graph layout metrics (i. e., testing construct validity with the differentiation by the "known groups" criterion [12]). To that end, we wanted to have 3 different groups of measures for each participant, so we produced 3 different visualizations with different levels of graph layout metrics.

P.1 Stimuli visualizations generation and metrics

We created node-link stimuli based on a transformed version of the "Les Misérables" dataset. This network was first created by Donald Knuth as part of the Stanford Graph Base. It contains 77 nodes corresponding

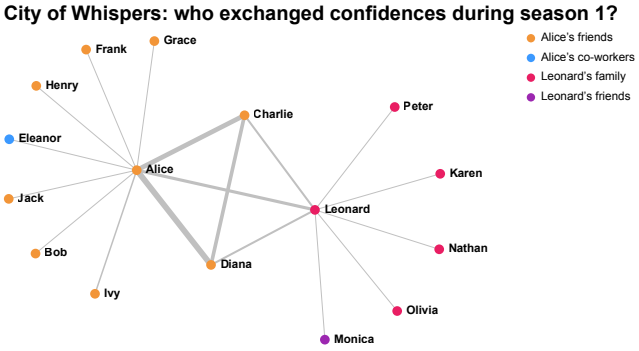

Fig. 73: Stimulus Ⓐ in our validation survey.

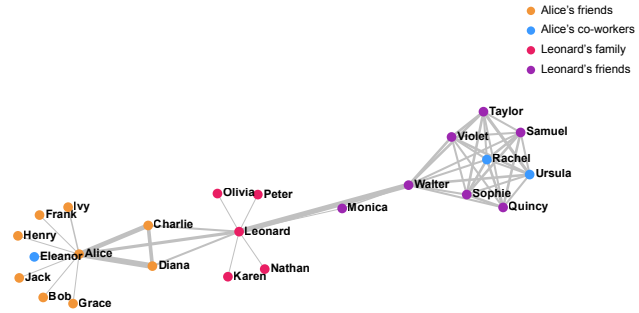

Fig. 74: Stimulus Ⓑ in our validation survey.

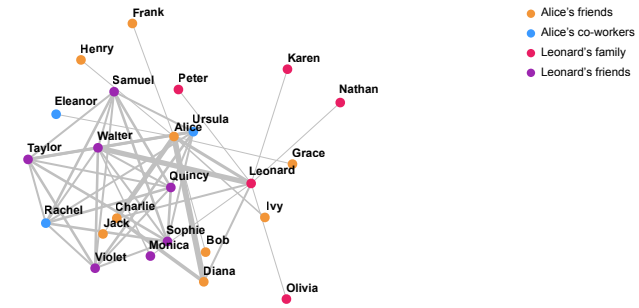

Fig. 75: Stimulus Ⓒ in our validation survey.

to characters of the novel, and 254 vertices connecting two characters whenever they appear in the same chapter. We retrieved the dataset from Mike Bostock's .block repository. We then replaced all characters' names by new, fictional first names of an imaginary TV show we called "City of whispers". Finally, we generated different visualizations using a simple force-directed graph generation script using the `D3.js` library (v6). We provide our code as supplemental material.

P.1.1 Graph layout metrics

As our goal was to compare the scale's ratings with graph aesthetics metrics, we implemented the `Greadability.js` library from R. Gove [43] (`github.com/rpgove/greadability`) within our code to calculate 4 layout metrics, which are described on the Greadability repository as follow:

- **Edge crossings:** "measures the fraction of edges that cross (intersect) out of an approximate maximum number that can cross."
- **Edge crossing angle:** "measures the mean deviation of edge crossing angles from the ideal edge crossing angle (70 degrees)."
- **Angular resolution (minimum):** "measures the mean deviation of adjacent incident edge angles from the ideal minimum angles

(360 degrees divided by the degree of that node)."

- **Angular resolution (deviation):** "measures the average deviation of angles between incident edges on each vertex."

To influence the layout, we use the same dataset but we vary three parameters: the number of plotted nodes and two "force" attributes in the `D3 force-directed` graph component. In particular, for the number of nodes we affected the following values: N = 16 nodes in df, and N = 24 nodes in B and C. For further details we refer the reader to our OSF Research log or to our stimuli generation code folder.

For A and B, we dragged elements to improve the clarity of the layout. We left C as it was. We then retrieved the calculated values from the `Greadability.js` library (which gets calculated again after each manipulation of the layout through user interaction and output in the console) and we save the generated SVG files, which we provide in our supplemental material folder.

We obtained the metrics described in Table 28. We noticed that all metrics corresponded to our presumed readability ranking of A > B > C except for the crossing angle where B > A. This metric is described in Gove's work [43] as a score on the average angle value of crossing between edges for non-adjacent nodes. The underlying assumption is that a 70° crossing is the ideal crossing angle for readability. However, stimulus A (Fig. 73) has only one such edge crossing, with an approximate value of 90°, hence its score is lower than B.

Table 28: Metrics obtained with the Greadability library for each node-link stimulus in our validation survey.

	Crossing	Crossing angle	Angular resolution min	Angular resolution dev	Greadability average
A	0.973	0.715	0.893	0.908	0.872
B	0.879	0.846	0.641	0.776	0.786
C	0.643	0.784	0.575	0.713	0.679

Finally, we modified the obtained SVG files in Adobe Illustrator to (1) modify nodes colors and add legends, (2) add titles, and (3) improve image fit in our survey screen—i. e., we rotated the full graph ***without altering the node-link layout***, ensuring continued validity of all graph layout metrics. We also provide the AI files in our supplemental material folder. Fig. 73 to Fig. 75 show the resulting images we used in our survey.

P.1.2 Color impairment simulations

Since we used color to encode groups of nodes in the visualizations, we chose color that would remain distinguishable for people with color vision deficiencies. We assessed the results in Adobe Illustrator using the View > Proof Setup for deuteranopia and protanopia, as shown in Fig. 76.

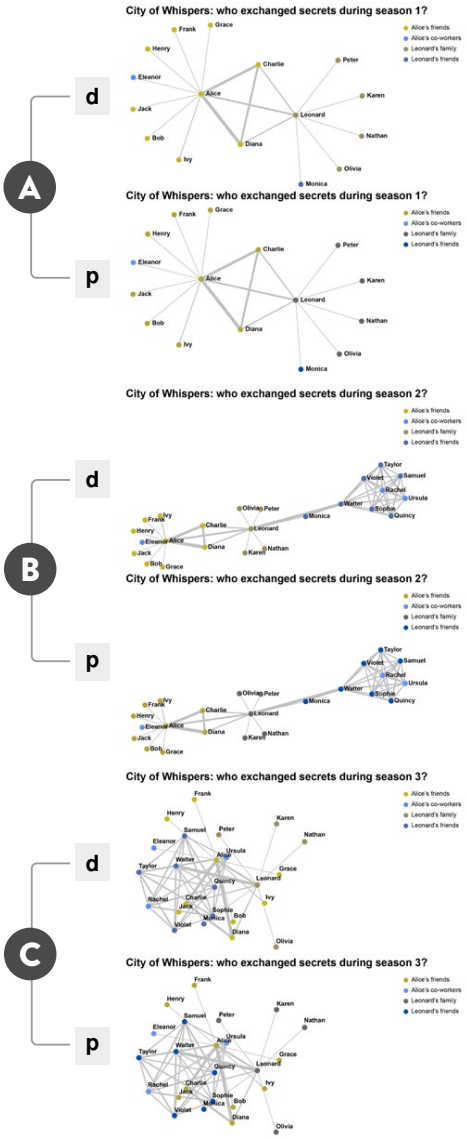

Fig. 76: Deuteranipa (d) and protanopia (p) simulations in Adobe Illustrator for our validation survey stimuli.

P.2 Reading tasks

Topology tasks are specific to node-link networks and were not part of our original analysis for reading tasks in the exploratory survey described in Sect. 5.1. To decide on reading tasks for the validation survey, we referred to Ghoniem *et al.*'s [40] work on assessing graph readability. We provide details on our selection of tasks in our OSF Research log.

As a result, we selected two types of task, which we repeated for each stimulus visualization in our survey:

- **Find the node with the maximum number of adjacent nodes.** e. g., "*Who is more popular?*". Using Amar *et al.*'s taxonomy [3] as a base to describe low-level visual analytics components, this task could be called a *Find Extremum Node* task. In our survey, we coded it as Find Extremum Node—`TaskFEN`.

  Because we used the same dataset for all stimuli, for each question we specified a subset of nodes in which the extremum had to be identified, which meant that the reader had to visually *Filter* the specified nodes as a preliminary reading step, before performing a *Determine range*, and finally what Amar *et al.* [3] call a *very low-level mathematical comparison task*.

- **Find the set of nodes adjacent to a node.** Instead of finding a set of nodes, we proposed a simplified version of this task, which could be better described as "Test direct connection", e. g."*Is A directly connected to B?*". In Amar *et al.*'s taxonomy [3], this would relates to a series of *Find* tasks, for which a reader could adopt different strategies:

  - *Find* source Node + *Determine range* on Adjacent Nodes → *Find* target Node among Adjacent Nodes, or
  - *Find* source Node + *Find* target Node → *Correlate*.

  In our survey, we coded this task as Find Adjacent Node—`TaskFAN`.

Although above-mentioned tasks were combinations of low-level tasks, it was possible to facilitate the "Find" task by using pre-attentive encodings to help participants filter the layout, similar to what Ghoniem *et al.* [40] did in their experimental design. We would thus ensure as low as possible level for the task. Since we did not want to change the colors for each task, we used cluster colors and references to these

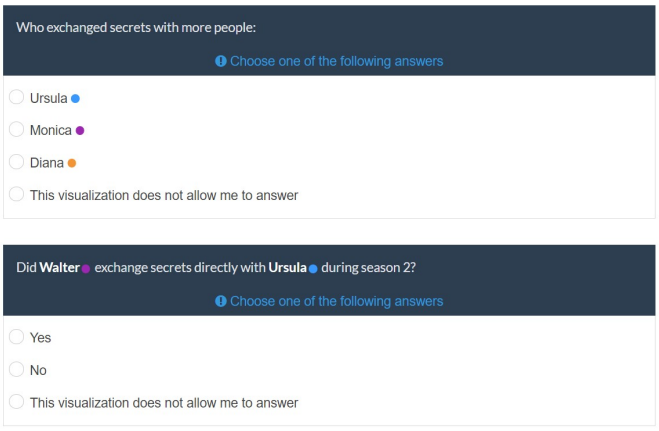

Fig. 77: Example screenshots of reading questions in our validation survey, in which we used color dots to facilitate the finding of relevant nodes in the visualization.

colors in the questions (as shown in Fig. 77). As a result, the *Find* tasks were not completely pre-attentive, but we facilitated them as much as possible.

Q Validation Survey: Additional Procedure Details

Here we provide additional details about the procedure and conditional design of our validation survey.

Our survey was structured in 4 main steps:

1. **Consent form.** Participants needed to read and approve of our consent form in order to proceed with the questionnaire. All subsequent questions were mandatory, with the exception of comments after ratings scales.
2. **Preliminary questions and instructions.** We then asked participants 1 question about color-vision deficiency before showing them an explanatory screen on how to read a node-link diagram such as the ones we used as stimuli visualizations. We asked participants to confirm their understanding of our explanations and then provided them with instructions regarding the experimental part of our survey.
3. **Survey stimuli reading and rating questions.** We assigned a random order of appearance to each stimulus. We report the resulting distribution of appearance orders further below in Table 29, and technical explanations in our OSF Research log. For each stimulus, participants aswered:
   (a) **All reading task questions** (see Sect. P.2). In addition, for the **first** visualization in order of appearance only, we asked a comprehension check question ("*What was the title of the visualization you just saw?*"). In this question we hid the node-link visualization. In line with Prolific attention and comprehension check policy, we gave participants two opportunities to answer correctly and they had the option of going back to the previous screen to read again the visualization and its accompanying text. Incorrectly answering this question twice was an exclusion criteria (similar to our exploratory survey).
   (b) **PREVis questionnaire.** For each subscale ◆◆◆◆, we displayed all rating items in a single screen with labeled 7-point Likert scale answer options as shown in Fig. 72. Subscales appeared in a random order, and items within subscales were also randomized.
   For each stimulus we also included one Instructional Manipulation Check (IMC, e. g., "For attention check purposes, please select **slightly agree** with this item"). We could not randomize the placement of such items, but we scattered them across different subscales and also modified the answer option we asked respondents to select—avoiding extreme values because, as we reported in our OSF Research log, participants from our pre-test interviews (exploratory survey) were uncomfortable in using extreme points of the Likert scale. Answering incorrectly more than once to IMCs was an exclusion criterion.
4. **Extraversion personality trait questions.** As explained in Sect. 6.1, in order to produce our Multi-Trait Multi-Method matrix (Fig. 4), we decided to compare PREVis ◆◆◆◆ scores with a measure of extraversion as personality trait. We used the 2 scale items related to extraversion from a validated 10-items version of the Big Five personality Inventory [84]: "I see myself as **extraverted, enthusiastic.**" and "I see myself as reserved, quiet." (for the latter we would reverse the rating before calculating average extraversion scores for participants).
   We presented these two items using the same answer options as our PREVis rating items and we specified that these questions were calibration items, unrelated to the visualizations participants had seen before.

R Validation Study Results

R.1 Stimuli randomization order distribution:a

We used the LimeSurvey platform to distribute our survey; LimeSurvey provides different levels of randomization functions, on which we do not have control in terms of parameters. We monitored the randomization order and noticed that stimuli orders of appearance were not equally distributed among participants. As such, the most common order was B, then C, then A.

Table 29: Order in which participants saw 3 node-link stimuli visualizations in our validation study.

	Count in A	Count in B	Count in C
Appeared 1st	34	65	49
Appeared 2d	53	38	57
Appeared 3d	61	45	42

R.2 Dimensionality, reliability and construct validity tests

We thoroughly validated PREVis by conducting different tests of dimensionality, reliability, and construct validity. We report in the main paper the key findings; in this section we describe the methods and provide some more details on our results. We provide all code and data to conduct these analyses in our supplemental material folder.

R.2.1 Tests of dimensionality

To validate the 4-factors dimensions, we followed the same steps as in the exploratory approach. We started with a parallel analysis (which we explained in Appx. J). From visual analysis on the scree plot in Fig. 78, we confirmed that 4 factors were appropriate to explain the variance in our data.

We then conducted Multi-Group Confirmatory Analysis (MG-CFA) (see Appx. L). Table 30 shows the model fit metrics, which allowed us to validate our 4-factor structure. We confirmed goodness of fit of our 4-factors structure with the following metrics: the Tucker–Lewis Index (TLI) was 0.97, the Comparative Fit Index (CFI) was 0.98, the Standardized Root Mean square Residual (SRMR) was 0.034. These metrics were in line with cutoff values recommended from Hu and Bentler [50]: fit indices such as TLI or CFI should be higher than .95, and SRMR should be lower than 0.9.

R.2.2 Tests of reliability

We conducted a first rough assessment of reliability by plotting the repeated measures correlation matrix for our collected data in Fig. 79, using the `rmcorr` package in `R`. This matrix also reflected clusters of covariance associated with our 4-factors structure.

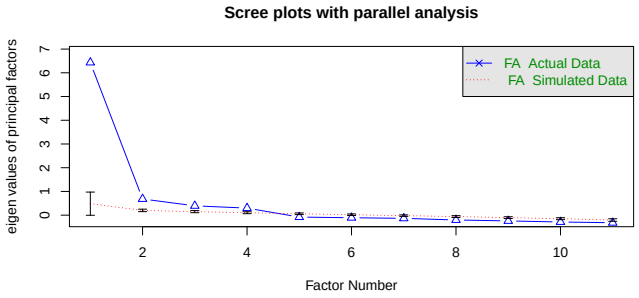

Fig. 78: Scree plot from parallel analysis with collected ratings from our validation survey.

Table 30: PREVis fit metrics from our validation study.

Fit indices	Full survey	Stimulus A	Stimulus B	Stimulus C
chisq	203	85	57	59
pvalue	0.000	0.000	0.023	0.013
cfi	0.979	0.966	0.987	0.984
tli	0.970	0.951	0.981	0.977
srmr	0.034	0.038	0.033	0.031
rmsea	0.073	0.092	0.059	0.062

Then, we tested the reliability of individual PREVis ◆◆◆◆ subscales by calculating Cronbach's alpha and McDonald's omega coefficient using `R's psych` package. Table 31 shows the reliability coefficients for each subscale, calculated for the full survey data and individually for each stimulus data subset. We obtained high reliability values for all subscales: we found the lowest value to be raw alpha = 0.877 for ◆ **LAYOUT** in Ⓒ. DeVellis and Thorpe consider such reliability values as "very good" [30].

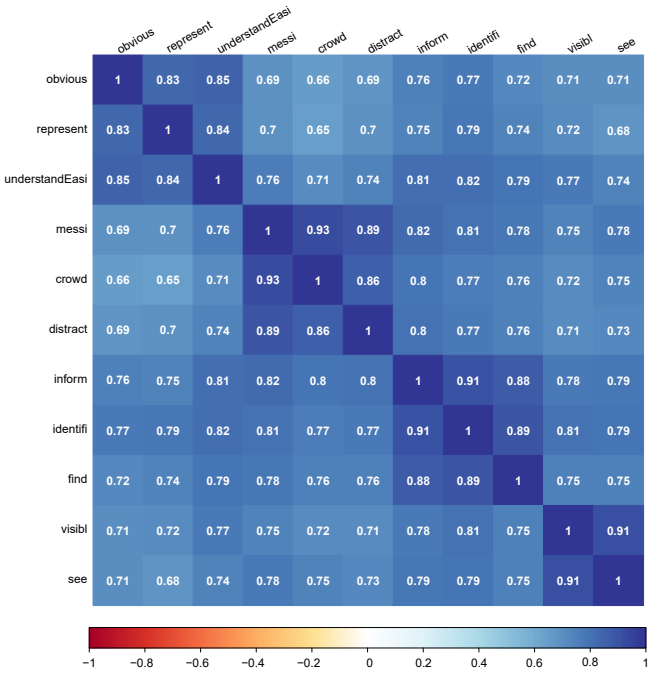

Fig. 79: Repeated measures correlation matrix PREVis ◆◆◆◆ items (see Table 32 for variable name correspondences with items statements). We generated this matrix using the `rmcorr` package in `R`.

Table 31: PREVis subscales' Cronbach's alpha and McDonald's omega reliability coefficients from our validation study.

Data set	Coefficient	understand	layout	dataRead	dataFeat
Full survey	**raw alpha**	0.936	0.953	0.955	0.946
	std alpha	0.936	0.953	0.956	0.946
	omega tot	0.937	0.953	0.956	0.946
Stimulus Ⓐ	**raw alpha**	0.926	0.899	0.895	0.911
	std alpha	0.928	0.900	0.896	0.912
	omega tot	0.929	0.902	0.901	0.912
Stimulus Ⓑ	**raw alpha**	0.914	0.884	0.918	0.932
	std alpha	0.914	0.885	0.918	0.932
	omega tot	0.914	0.886	0.920	0.932
Stimulus Ⓒ	**raw alpha**	0.901	0.877	0.933	0.892
	std alpha	0.901	0.886	0.933	0.892
	omega tot	0.901	0.891	0.933	0.892

Table 32: Variable names correspondences with PREVis items in our analyses.

	Item code	Item statement
◆ UNDERSTAND	obvious	It is obvious for me how to read this visualization
	represent	I can easily understand how the data is represented in this visualization
	understandEasi	I can easily understand this visualization
◆ LAYOUT	messi	I don't find this visualization messy
	crowd	I don't find this visualization crowded
	distract	I don't find parts of the visualization distracting
◆ DATAFEAT	visibl	I find data features (for example, a minimum, or an outlier, or a trend) visible in this visualization
	see	I can clearly see data features (for example, a minimum, or an outlier, or a trend) in this visualization
◆ DATAREAD	inform	I can easily retrieve information from this visualization
	identifi	I can easily identify relevant information in this visualization
	find	I can easily find specific elements in this visualization

R.2.3 Tests of construct validity

We conducted multiple tests to confirm the validity of PREVis. Construct validity testing consist in verifying that the instrument correctly targets the construct that it was built to measure: in our case, perceived readability in data visualizations. Following recommendations from our reference literature in scale development [12, 30], and as we state in Sect. 6.1, we designed our validation study to test for convergent and discriminant validity using the multi-traits multi-method (MTMM) approach:

1. **Inter-subscales reliability:** our subscales' scores should highly and positively correlate among themselves, indicating the existence of a shared latent variable in respondents: the *perceived readability* construct.
2. **Discriminant validity:** our subscales' scores should not correlate with extraversion scores measured with a similar method (i. e., in our, case, a 7-point Likert scale from the Big Five personality Inventory short version [84]) because it is a different, unrelated construct.
3. **Convergent validity:** our subscales' scores should positively correlate with Greadability metrics [43] because they are also indicators related to readability, but measured using a different method.

A MTMM correlation matrix allows us to check all of these criteria at once. We first generated MTMM matrices in R using the `corr` function in the `psych` package, which assumes independence of observations. It is a conservative way to analyze our results because, by assuming independence of all PREVis measurements, between-participant differences add error to the measures, possibly reducing their covariance across stimuli. Our study's design, however, also allowed for within-participant assessment of 1 inter-subscales reliability and 3 convergent validity using repeated measures correlations:

1. **Inter-subscales reliability:** each of our 148 participants rated each stimulus (A, B, C), therefore we collected 3×148 measures for each ◆◆◆◆ subscale. This allowed us to conduct a repeated measures correlation analysis (i. e., correlations based on within-participant covariance of PREVis ratings across stimuli).
2. **Discriminant validity:** because personality traits are stable, we measured extraversion only once per participant, thus collecting 148 observations. Because extraversion scores did not vary across experimental conditions, conducting a repeated measures correlation analysis caused computational errors. These errors stemmed from the underlying model encountering near-zero values during the covariance analysis, leading to numerical precision issues and falsely negative sums of squares.
3. **Convergent validity:** we calculated Greability measures for each stimulus (see Table 28) and copied these values in the result table for each participant, therefore we have 3×148 values. This allowed us to conduct **a repeated measures correlation analysis** (i. e., correlations based on within-participant covariance to remove between-participant variance).

Scale-level correlation matrices. Fig. 80 shows a scale-level version of the independent measures MTMM matrix generated with the `cov2cor` function in the `stat` package in R. This conservative approach confirms all 3 criteria despite the between-participant noise: inter-subscales reliability (positive and high correlations in 1), discriminant validity (correlations close to 0 in 2), and convergent validity (positive correlations in 3).

Fig. 81 shows a scale-level version of the repeated measures MTMM matrix. Eliminating between-subject variance allows to find much higher correlations between PREVis ◆◆◆◆ ratings and Greadability metrics, strengthening convergent validity of our instrument. Inter-subscales reliability also slightly improves. As extraversion was not collected with repeated measures, this tool does not allow us to calculate discriminant validity. While attempting to process the data, the `rmcorr` package in R returned NA values for correlations between ◆◆◆◆

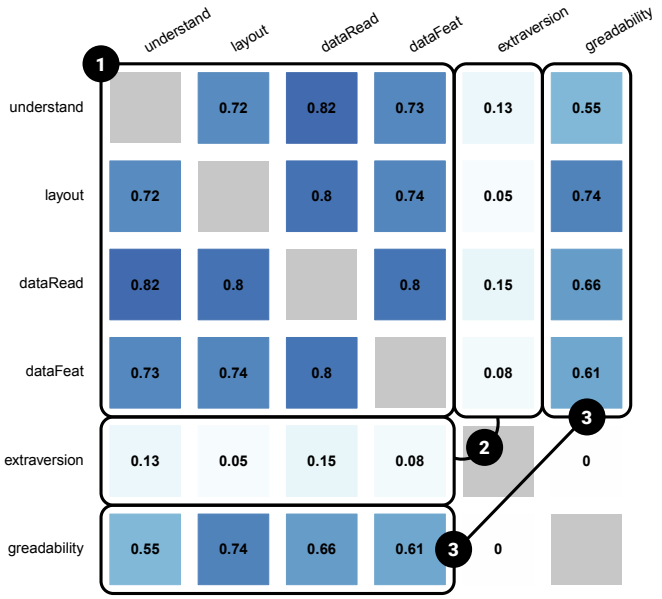

Fig. 80: Multi-trait multi-method (MTMM) **independent measures** correlation matrix at the scale level: 1 reliability among PREVis subscales, 2 discriminant validity from an unrelated personality trait in respondents, and 3 convergent validity with graph layout metrics.

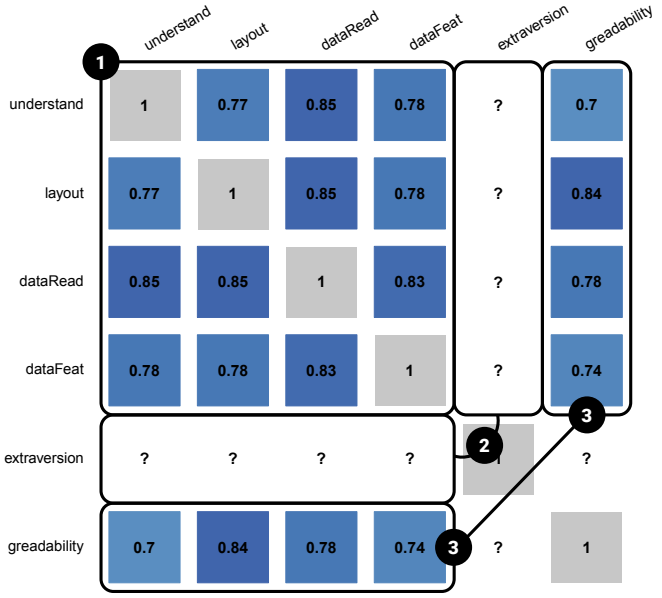

Fig. 81: Multi-trait multi-method (MTMM) **repeated measures** correlation matrix at the scale level: 1 reliability among PREVis subscales, 2 discriminant validity from an unrelated personality trait in respondents, and 3 convergent validity with graph layout metrics.

subscales and Greadability metrics, as well as for correlations between Greadability metrics and extraversion ratings.

This is explained technically because Greadability has fixed values across participants within each condition (stimuli A, B, and C), and extraversion ratings have a fixed values within each participant across the three conditions. This can result in very small or zero variance within those groups of variables. When the variance is extremely small, numerical precision issues can lead to negative sum of squares values, which leads to an invalid calculation for the repeated measures correlation coefficient. We document the problem and the code we used to apply a selective adjustment based on a small tolerance value of 1e−10 it in our OSF Research log.

Item-level correlation matrices. Fig. 82 shows the independent measures MTMM matrix at the item-level. The first 11 lines and rows correspond to PREVis items, grouped by subscale, and show high correlations; the next 2 elements correspond to the "extraversion" items from the BFI 10-items scale [84] and show an absence of correlation with all data collected in this survey; the last 4 rows and columns correspond to the `Greadability.js` metrics [43] from Table 28. In this part of the item-level MTMM, one of the items in the set of `Greadability` metrics shows negative correlation to the others. It corresponds to the **"`crossingAngle`"** metric. We discuss this metric above in Sect. P.1.1. Fig. 83 shows the repeated measures MTMM matrix at the item-level.

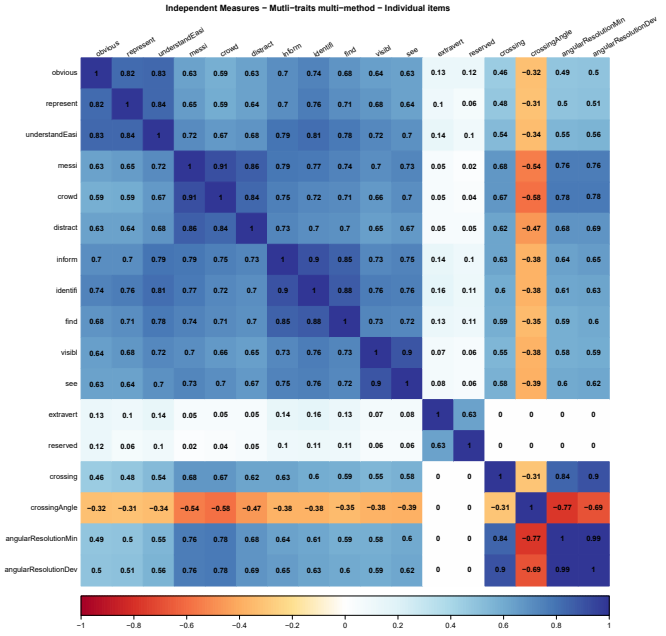

Fig. 82: Item-level Multi-Trait Multi-Method **independent measures** correlation matrix from our validation study.

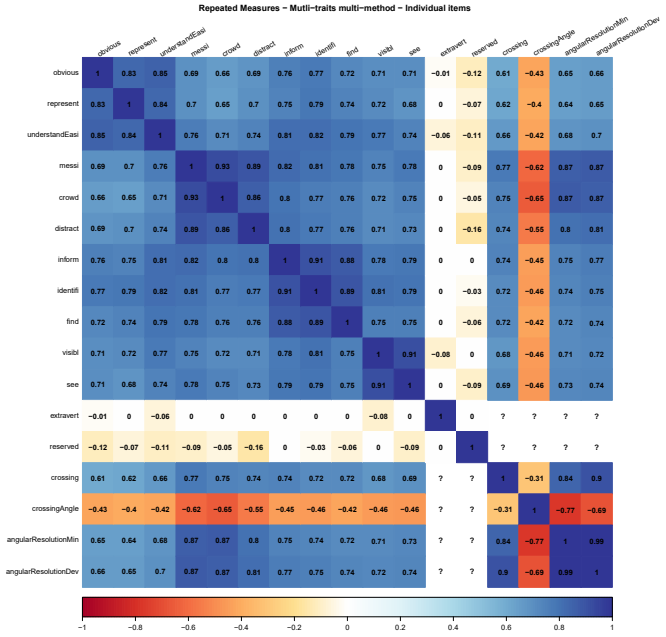

Fig. 83: Item-level Multi-Trait Multi-Method **repeated measures** correlation matrix from our validation study.

We share in Fig. 4 in the main paper a scale-level composite MTMM matrix where correlations across stimuli for PREVis and Greadability are calculated with the repeated measure approach from the `rmcorr` package in `R`, and the correlations for the *extraversion* between-participants variable is calculated based on the `cov2cor` function in `stat` package from `R`.

Finally, we plotted PREVis ◆◆◆◆ subscales' scores with 95% CI across stimuli to test their validity with what Boateng *et al.* call "*differentiation by known groups*" [12]: (Fig. 84 to Fig. 87). We verified that we could distinguish between the 3 node-link, in the expected order: Ⓐ > Ⓑ > Ⓒ.

We also plotted individual items ratings for each ◆◆◆◆ subscale and stimulus in Fig. 88 to Fig. 99.

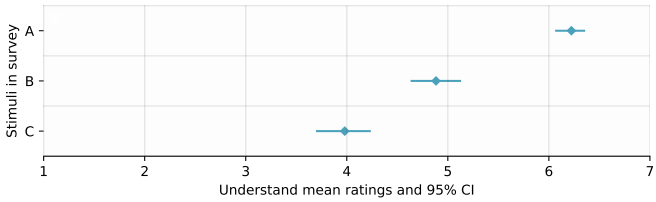

Fig. 84: ◆ **UNDERSTAND** scores across stimuli in our validation study allow to distinguish and correctly rank Ⓐ > Ⓑ > Ⓒ.

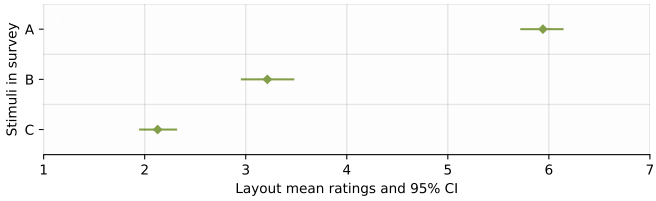

Fig. 85: ◆ **LAYOUT** scores across stimuli in our validation study allow to distinguish and correctly rank Ⓐ > Ⓑ > Ⓒ.

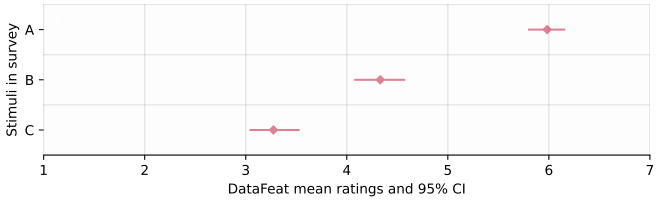

Fig. 86: ◆ **DATAFEAT** scores across stimuli in our validation study allow to distinguish and correctly rank Ⓐ > Ⓑ > Ⓒ.

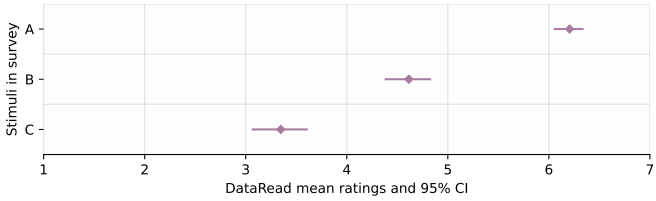

Fig. 87: ◆ **DATAREAD** scores across stimuli in our validation study allow to distinguish and correctly rank Ⓐ > Ⓑ > Ⓒ.

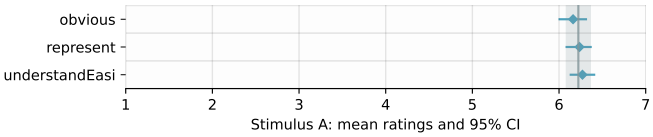

Fig. 88: ◆ **UNDERSTAND** individual items scores and average score in Ⓐ.

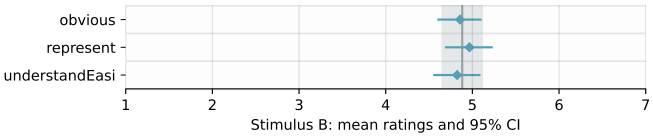

Fig. 89: ◆ **UNDERSTAND** individual items scores and average score in Ⓑ.

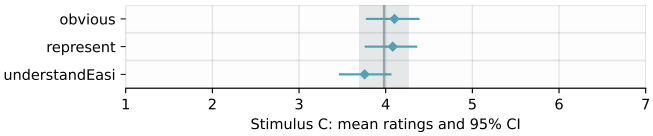

Fig. 90: ◆ **UNDERSTAND** individual items scores and average score in Ⓒ.

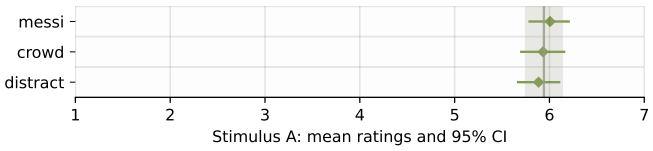

Fig. 91: ◆ **LAYOUT** individual items scores and average score in Ⓐ.

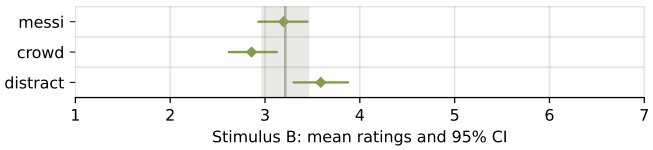

Fig. 92: ◆ **LAYOUT** individual items scores and average score in Ⓑ.

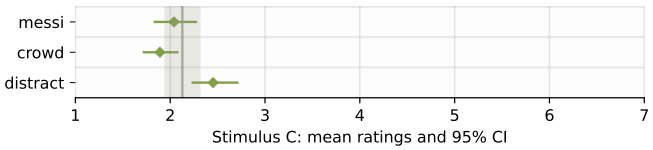

Fig. 93: ◆ **LAYOUT** individual items scores and average score in Ⓒ.

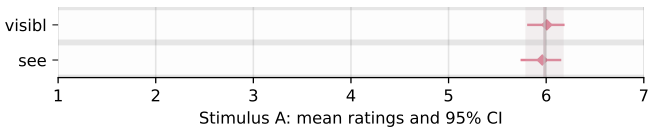

Fig. 94: ◆ **DATAFEAT** individual items scores and average score in Ⓐ.

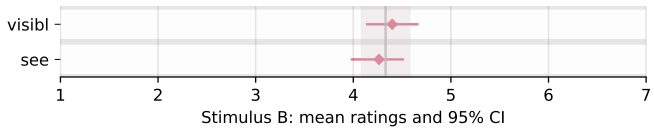

Fig. 95: ◆ **DATAFEAT** individual items scores and average score in Ⓑ.

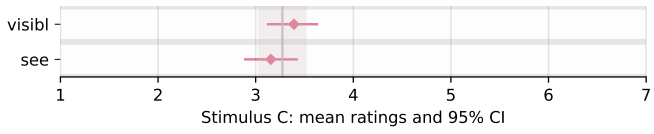

Fig. 96: ◆ **DATAFEAT** individual items scores and average score in Ⓒ.

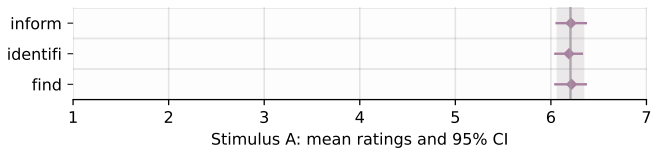

Fig. 97: ◆ **DATAREAD** individual items scores and average score in Ⓐ.

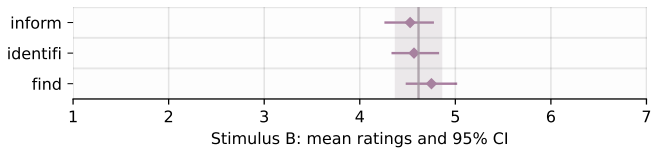

Fig. 98: ◆ **DATAREAD** individual items scores and average score in Ⓑ.

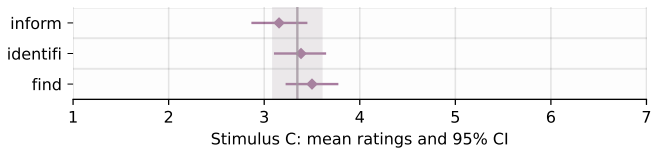

Fig. 99: ◆ **DATAREAD** individual items scores and average score in Ⓒ.

FIGURE CREDITS AND COPYRIGHT

Fig. 11 is © 2010 IEEE and we reused it (with permission) from Bezerianos *et al.*'s paper [11]. All the remaining figures in this appendix are our own, and for them we retain the copyright but allow them to be used here. They are available under the Creative Commons CC BY 4.0 license and we share them in our supplemental material folder at osf.io/9cg8j.