PREVis

Perceived Readability Evaluation in Visualization

Questionnaire items and implementation recommendations

Understand subscale

`obvious`	It is **obvious** for me how to read this visualization
`represent`	I can easily understand **how the data is represented** in this visualization
`understand`	I can **easily understand** this visualization

Layout subscale

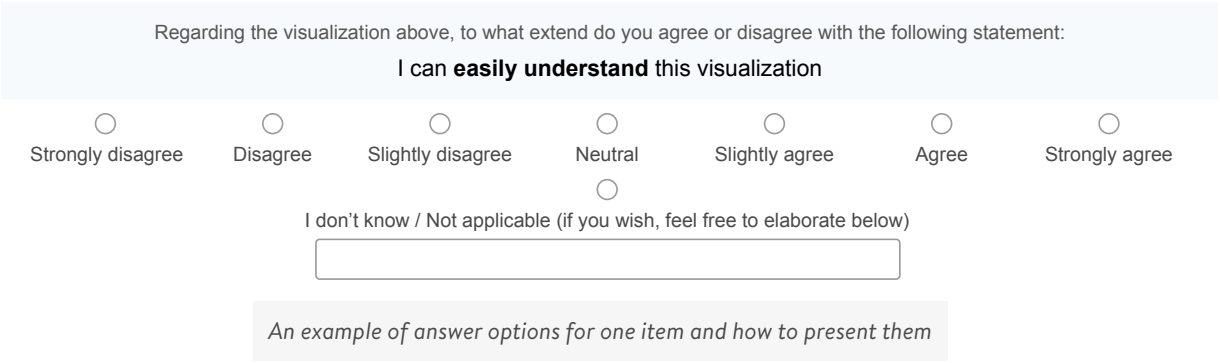
`messy`	I **don't** find this visualization **messy**
`crowd`	I **don't** find this visualization **crowded**
`distract`	I **don't** find **distracting parts** in this visualization

DataRead subscale

`find`	I can easily **find specific elements** in this visualization
`identify`	I can easily **identify relevant information** in this visualization
`information`	I can easily **retrieve information** from this visualization

DataFeat subscale

`visible`	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



An example of answer options for one item and how to present them

HOW TO USE PREVis?

In this document, we provide practical considerations for effective use, implementation and analysis of PREVis measurements. We gathered the following content to guide researchers:

FAQ

Study protocols examples

PDF forms for PREVis

FAQ

Is PREVis appropriate for my specific visualization?

We developped and tested PREVis to help provide insight on readability for many different types* of visualizations. It is important to note, however, that we developed PREVis with a focus on readability in static images, and **NOT** to evaluate a system's **interactive features**—even features that are targeted at improving readability of visualizations. If you would like to use PreVis on visualizations in an interactive tool where visualizations can change, you can still use PreVis but you can only test specific instances of the visualizations that your tool can produce. PreVis can in this case, for example, help you to compare which representations types might be more or less readable for certain dataset sizes, dataset types or using which color scales, label types, etc.

Similarly, PREVis might **lack some dimensions** to measure readbility **3D environments** such as virtual reality, or physical visualizations in the real world. Such contexts would ideally require additional work prior to using PREVis. As a first step, you should conduct a qualitative study with users and experts to assess the need for expanding or adapting PREVis items to your specific context. Depending on the results, you might then need to generate new items and conduct an exploratory study followed with a validation study similar to the one we ran for PREVis; or it's possible that adapting PREVis items and running a validity study would be enough.

How many participants do I need?

We cannot answer this question a-priori because it depends on what you want to do with the results. If you want to measure a specific difference in the perceived readability between two visualizations (see Study 1 below), you could conduct a power analysis. If you want to use the PREVis results to initiate a discussion with participants (see Study 2 below), you can choose the number of participants by following a principle of saturation as in qualitative research.

Is PREVis appropriate for my specific participants?

We developed and tested PREVis with a general population. We went through a dedicated qualitative research phase called « cognitive interviews » where we showed the scale to people who were not visualization researchers/experts to ensure that the items of the scale would be understandable to a general population. Our final validation study was also conducted with a general population filling out PREVis. We are, therefore, confident that the scale can be used by most study participants. However, we did not test the scale with children or people with any kinds of cognitive or other impairments and cannot make any recommendations about the use of the scale with such study populations.

When do I deploy PREVis?

This depends on the type of study you run. Generally, we recommend deploying PREVis shortly after people use the visualization you are studying.

In which form do I deploy PREVis?

You can deploy PREVis on paper (see the last pages in this document) or by creating a digital version. We will soon share example forms for different survey platforms.

Are the 4 subscales independent in PREVis ◆ ◆ ◆ ◆?

No, the 4 dimensions of our instrument are not entirely independent: together, they indicate how easily people feel they can read a data visualization—which we call the "perceived readability" construct. We have observed that all subscales (i.e., dimensions) share common variance in respondents; however, there are also important differences which require researchers to analyze each subscale separately. As such, **you should not aggregate PREVis ◆ ◆ ◆ ◆ subscales' individual ratings into an average score** (see next question: how do I analyze the ratings?).

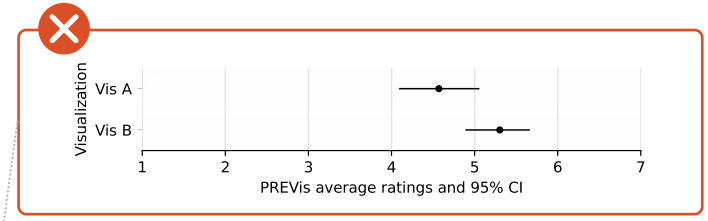
It also means that, if you are only interested in studying a specific component of perceived readability, you can drop subscales that are not relevant. For example, if you are testing a system where people cannot read individual data points but only trends or clusters, you might drop the ◆ **DataRead** subscale; or, if you are interested in studying the layout clarity among experts that are already trained in reading the type of visualization you are using, you could drop the ◆ **Understand** subscale.

It's worth noting that the ◆ **Understand** and the ◆ **Layout** subscales might target two formative dimensions of perceived readability: if a person feels that they do not know how to read a visualization or find the layout too cluttered, this would contribute to forming a lower perceived readability for that person. In contrast, the ◆ **DataRead** and ◆ **DataFeat** scales might target reflective dimensions: poor perceived readability of a visualization will be reflected in participants experiencing difficulties in their attempts to read data or data features. These hypothetical properties of our four subscales are not yet established and require further studies; however, the formative or reflective nature of these indicators does not affect their validity for measuring perceived readability in studies.

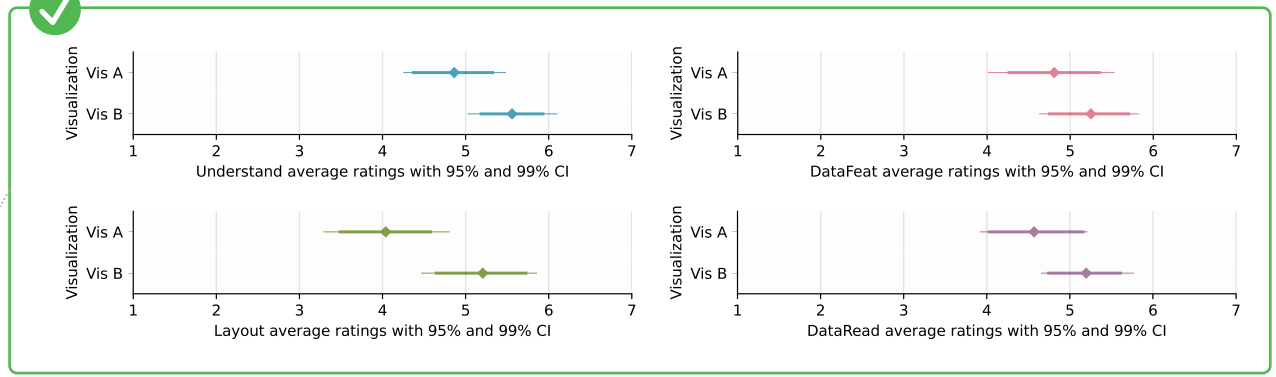
*While it is virtually impossible to test PREVis on all existing types of visualizations, along the development process we used a bar chart, a histogram, two line charts, a pie chart, a scatterplot, a bubble chart, a choropleth map, node-link representations, and a novel type of representation for genealogical trees called GeneaQuilts (see Bezerianos et al., 2010).

How do I analyze the ratings?

You calculate an average (mean) for each subscale. Do NOT calculate a mean for the entire instrument. There might be small differences for a specific perceived readability dimension while others have larger differences. This will only be captured by an individual subscale: by calculating an average score, you might flatten these differences and it will hinder your analysis.



Avoid calculating an average PREVis score like the figure to the left; instead, we recommend you plot individual subscales' averages individually like the plots below.



The example data to the right was collected during an independent study with two visualizations and 34 participants. If plotted together, the average PREVis ratings do not differ much (about 0.75 difference between point estimates).

Upon closer look, however, ◆ **Layout** ratings showed a higher difference between visualizations A and B, while the ◆ **DataFeat** readability ratings only showed a very small difference. We can only make such observations by analyzing each subscale separately.

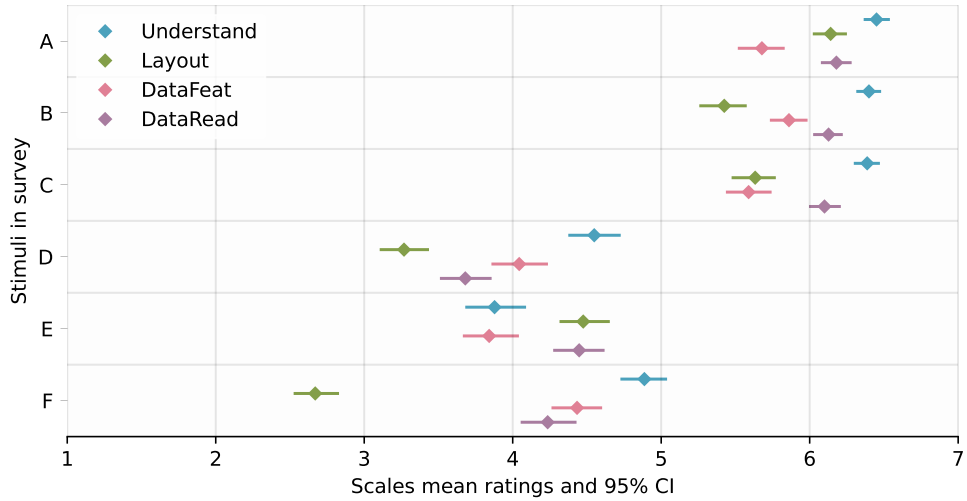
Visualizations may also have different ranks depending on the subscale. This is illustrated by looking at ratings for visualizations D, E and F in our exploratory survey in the adjacent figure.

While visualization E obtained higher ◆ **Layout** ratings than visualizations D and F, the situation was reversed for ◆ **Understand**. These ratings mean that participants found it more difficult to understand how to read visualization E, even though its layout was visually clearer than visualizations F and D. Flattening PREVis score would hinder the possibility to identify what the readability issues are.

We provide example plots, `csv` files and a `Python` notebook to generate the plots in the osf.io/9cg8j/ folder `/Using PREVis - plots`.

Below: ratings from our between-participants exploratory study described in the PREVis paper (see Fig. 29 in Appendix O in the full PREVis paper).

When looking at ratings for visualizations D, E and F, we observe different ranks for the Understand or the Layout subscales' ratings.



STUDY 1: VISUALIZATION A VS. VISUALIZATION B

SCENARIO

You have two visualizations you would like to compare. Maybe...

- you developed Visualization A and want to compare it to the state-of-the-art, which is Visualization B, or
- you developed both Visualization A and B and want to choose between them

STUDY DESCRIPTION

This is a within-subjects (lab or online) study in which participants will perform one or multiple tasks with each visualization. Your goal is to study specific objective measures, perhaps error rates and time, but also potentially more subjective measures related to aesthetics, preference, required effort, etc. You want to deploy PreVIS to also study each visualization's perceived readability.

HYPOTHESES

You will have multiple hypotheses for your study depending on the metrics you include. For readability, a possible hypothesis could be:

H1: Participants perceive Visualization B to be more readable than Visualization A because [it uses an encoding for quantity that is considered more effective| it is rendered at a larger size | past studies have shown Visualization B to be superior to Visualization A in a similar scenario | ...]

OBJECTIVES

The study's primary objectives do not have to be framed around readability. They could be framed around measuring performance, cognitive load, or other metrics, including readability. In our scenario measuring readability is a secondary objective to help to potentially explain your results. For example, you could have the following objectives:

Primary Objective Example

Evaluate participants' task performance with Visualization A and Visualization B for Tasks T1, T2, and T3.

Secondary Objective Example

Compare participants' perceived readability following the performance of T1, T2, and T3.

Alternative Primary Objective Example

Evaluate participants' task performance with Visualization A and Visualization B for Tasks T1, T2, T3. Visualization B is a variant of Visualization A that just differs in layout (e.g. Vis A is a network visualization using layout Algorithm 1, and Vis B uses Algorithm 2). You want to find out if people perform better with Algorithm 1 than Algorithm 2.

Alternative Secondary Objective Example

Compare participants perception of the visual clarity of the layout produced with Algorithm 1 (Vis A) and Algorithm 2 (Vis B). In this case you would deploy PREVis only using the "Layout" subscale.

COLLECTED DATA

- Measurements from your primary objective
- PREVis scores per participant and item (grouped by subscale)
- Potentially responses to PREVis in case participants did not know how to answer

STUDY POPULATION

The number of participants you need depends on your study design and primary objective.

STUDY PROCEDURE

In our scenario an ideal option is to start the experiment by briefly presenting Visualization A and B, although we acknowledge that this is not always possible.

In any case, participants will then use one of the two visualizations chosen randomly, perform the three tasks with the first visualization, and only then fill out PREVis. Next, they continue the study with the second visualization and fill out PREVis again.

We do not recommend to deploy PREVis before people have had to actually read each visualization.

We also do not recommend to ask people to fill out PREVis for both visualizations at the end of the study.

DATA ANALYSIS

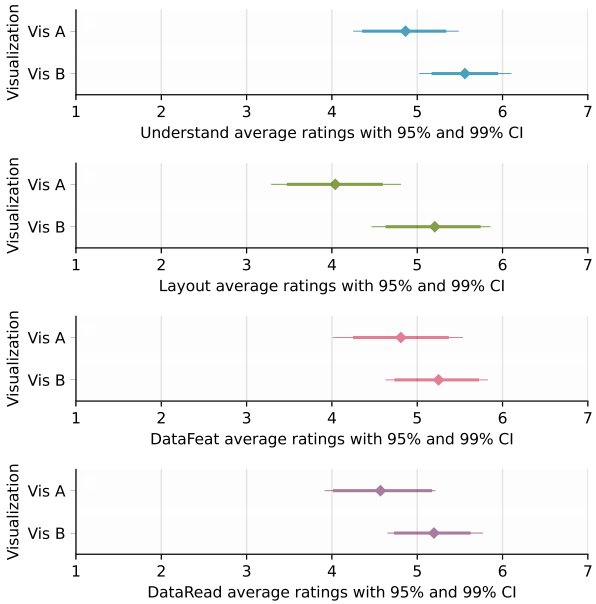
For PreVIS you calculate a mean for each subscale individually per participant. Next, you may perform inferential statistics using bootstrap confidence intervals to provide evidence of a difference between the scores for each visualization and subscale. See Dragicevic (2016) to get some help on how to do that (open access on hal.science/hal-01377894).

You can also refer to our data anlysis notebooks in the `/Using PREVis - plots` folder and/or validation study folder on osf.io/9cg8j/.

1. Point estimates with CI

You can plot point estimates for each subscale average with 95% and 99% bootstrap confidence interval (using, for example, the `pointplot` function with `matplotlib` and `seaborn` librairies in `Python`).

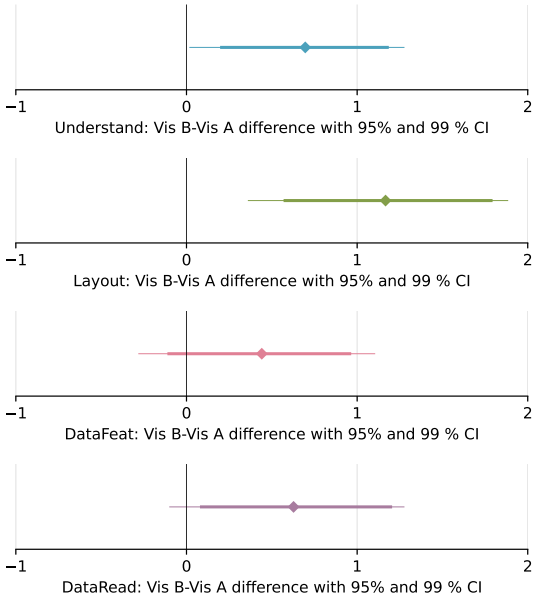
Below: Point estimate plots for each PREVis subscale with 95% and 99% confidence intervals



2. Repeated measures differences with CI

In this scenario, a repeated-measures difference analysis will allow you to further analyze the differences between visualizations A and B ratings. For each subscale and participant, you can calculate the difference between Vis A and Vis B, and then plot point estimates for these values with bootstrap confidence intervals. Or you could rely on existing methods for tests between two repeated measures, such as `ttest_rel` in the `scipy` library in `Python`.

Below: B-A difference point estimate for each PREVis subscale with 95% and 99% confidence intervals



INTERPRETATION OF THE RESULTS:

In this scenario you should report both analyses and focus on the size of the difference between the scores of each visualization. A difference between the scores provides evidence that participants found one to be more readable than the other. Remember that no single study can provide conclusive evidence.

Within-subjects designs aim to control for between-participant differences. Assessing the correlations between repeated measures (A and B ratings) for each subscale can help prevent misinterpretation of the confidence intervals. Positive correlations indicate consistency of measurements: it means that if Participant 1 gave a higher rating to Vis A than Participant 2, then Participant 1 will also tend to rate Vis B higher than Participant 2. Positive correlations in repeated measures can reduce margin of error for confidence interval estimation. Negative correlations, however, might indicate measurement error and increase the margin of error in the confidence interval estimation (Cumming and Finch, 2005). You can plot pairs of repeated measures on scatterplots with a regression line to help assess such correlations.

As PREVis measures “perceived” readability, the results will depend on each participant and study context. As such, we do not recommend to simply compare numbers from one experiment to another very different experiment.

STUDY 2: QUALITATIVE STUDY ABOUT A NEW VISUALIZATION

SCENARIO

You have developed a new visualization and would like to more deeply understand how people experience using it.

STUDY DESCRIPTION

This is a lab study in which participants will use your visualization, either freely or with tasks you give to them. Your goal is to better understand how you can further improve the visualization.

HYPOTHESES

A study such as this is exploratory and does not require hypotheses.

OBJECTIVES

The study's primary objectives is to explore how people experience your visualization, what works and what seems difficult. In our scenario measuring readability is one objective together with observations and responses to interview questions you will collect.

COLLECTED DATA

- Video + audio recordings of people using the visualization under a think-aloud protocol
- Video + audio recordings of an interview you conduct with the participants about how they experienced the visualizations
- Results from PREVis

STUDY POPULATION

The number of participants you need depends on your goal. If you want to conduct an in-depth qualitative study a common choice is to follow the principle of saturation.

STUDY PROCEDURE

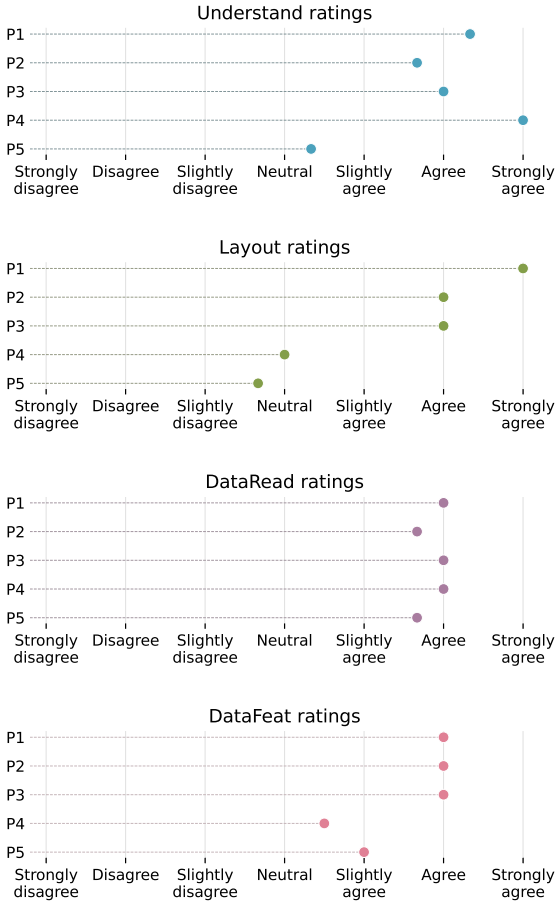
You should follow recommended study procedures for a think-aloud/observational usability or user experience study (again, depending on your goal). You can deploy PREVis at the end and then interview participants about how they answered the scale. Using PREVis in this way, it may serve as a conversation starter to better understand the different factors related to perceived readability. You could also fill out PREVis together with the participants by asking them the individual scale question and collect explanations at the same time.

To note: we have not experimented with using PREVis this way and as such, our recommendations stem from using questionnaires in qualitative research more broadly.

DATA ANALYSIS

You may still analyze PREVis the same way as in Study 1, analysis 2 (that is, using only the point estimate analysis with confidence intervals), but depending on the number of participants you could also simply choose the plot average scores for each scale and person, or a mean with a standard deviation.

You will also collect rich qualitative responses on participants' choices you can analyze using your favorite qualitative data analysis methods.



Example plots for 5 participants in a qualitative study.

PREVis

Printable questionnaires

SIMPLE QUESTIONNAIRE (1 page)

7-points Likert answer options and one comment field for each subscale (similar to our design in the validation study)

SEE PRINTABLE VERSION p. 8

DETAILED QUESTIONNAIRE (2 pages)

7-points Likert answer options + “I don’t know / Not applicable” (to treat as NA), with an additional comment field for NA answers. This comment fied should not be mandotry for participants to fill, but researchers should provide it as a separate paper sheet to allow respondents who wish so to comment directly as they answer the questionnaire.

SEE PRINTABLE VERSION pp. 9-10

PARTICIPANT ID:

VISUALIZATION ID:

	Strongly disagree	Disagree	Slightly disagree	Neutral	Slightly agree	Agree	Strongly agree	Optional comment
It is **obvious** for me how to read this visualization	○	○	○	○	○	○	○	
I can easily understand **how the data is represented** in this visualization	○	○	○	○	○	○	○	
I can **easily understand** this visualization	○	○	○	○	○	○	○	
I **don't** find this visualization **messy**	○	○	○	○	○	○	○	
I **don't** find this visualization **crowded**	○	○	○	○	○	○	○	
I **don't** find **distracting parts** in this visualization	○	○	○	○	○	○	○	
I find data features (for example, a minimum, or an outlier, or a trend) **visible** in this visualization	○	○	○	○	○	○	○	
I can **clearly see** data features (for example, a minimum, or an outlier, or a trend) in this visualization	○	○	○	○	○	○	○	
I can easily **find specific elements** in this visualization	○	○	○	○	○	○	○	
I can easily **identify relevant information** in this visualization	○	○	○	○	○	○	○	
I can easily **retrieve information** from this visualization	○	○	○	○	○	○	○	

PARTICIPANT ID:

VISUALIZATION ID:

	Strongly disagree	Disagree	Slightly disagree	Neutral	Slightly agree	Agree	Strongly agree	I don't know / Not applicable
It is **obvious** for me how to read this visualization	○	○	○	○	○	○	○	○
I can easily understand **how the data is represented** in this visualization	○	○	○	○	○	○	○	○
I can **easily understand** this visualization	○	○	○	○	○	○	○	○
I **don't** find this visualization **messy**	○	○	○	○	○	○	○	○
I **don't** find this visualization **crowded**	○	○	○	○	○	○	○	○
I **don't** find **distracting parts** in this visualization	○	○	○	○	○	○	○	○
I find data features (for example, a minimum, or an outlier, or a trend) **visible** in this visualization	○	○	○	○	○	○	○	○
I can **clearly see** data features (for example, a minimum, or an outlier, or a trend) in this visualization	○	○	○	○	○	○	○	○
I can easily **find specific elements** in this visualization	○	○	○	○	○	○	○	○
I can easily **identify relevant information** in this visualization	○	○	○	○	○	○	○	○
I can easily **retrieve information** from this visualization	○	○	○	○	○	○	○	○

PARTICIPANT ID:

VISUALIZATION ID:

	I don't know / Not applicable: feel free to elaborate below
It is **obvious** for me how to read this visualization	
I can easily understand **how the data is represented** in this visualization	
I can **easily understand** this visualization	
I **don't** find this visualization **messy**	
I **don't** find this visualization **crowded**	
I **don't** find **distracting parts** in this visualization	
I find data features (for example, a minimum, or an outlier, or a trend) **visible** in this visualization	
I can **clearly see** data features (for example, a minimum, or an outlier, or a trend) in this visualization	
I can easily **find specific elements** in this visualization	
I can easily **identify relevant information** in this visualization	
I can easily **retrieve information** from this visualization	