UNIVERSITY OF COLOGNE

DOCTORAL THESIS

# Visual Communication of Large Value Ranges and Visual Validation of Regression Models


*Author:*
**Daniel Braun**

*Supervisor:*
Prof. Dr. Tatiana von Landesberger

*Examiner:*
Prof. Dr. Manuela Waldner

*Examiner:*
Prof. Dr. Natalia Andrienko


*Dissertation for the degree of*
*Doctor of Natural Sciences (Dr. rer. nat.)*
*in the*
Faculty of Mathematics and Natural Sciences



Date of submission: July 7, 2025
Date of the oral exam: October 6, 2025

*"You miss 100 percent of the shots you don't take."*

Wayne Gretzky

# *Abstract*

This dissertation investigates two underexplored challenges in visual data analysis: the effective communication of large value ranges and the visual validation of statistical (regression) models. Within the visual analytics framework, both challenges target critical parts in the data analysis process — namely, visual encoding and model building — in which visualization plays a pivotal role in enabling human insights.

The first topic focuses on the visualization of large value ranges, especially in time-dependent data. Data with large value ranges are data sets whose values span several orders of magnitude. Standard visualizations, such as linear or logarithmic scales, often fail to support readability or accurate comparison across orders of magnitude. To address this, this dissertation proposes novel visualization techniques that explicitly encode both mantissa and exponent components through refined visual mappings, including a nested color scheme, a scale that bridges the strengths of linear and logarithmic axes, and multiple visual designs for single and multiple time-series data. Empirical user studies across a range of tasks — such as identification, discrimination, and estimation — demonstrate that these techniques significantly improve task accuracy, response time, and confidence in interpretation across domain-agnostic data sets. The contributions extend beyond nominal data to support complex time-series structures, for which large value ranges are common across scientific and public domains, introducing scalable designs that enhance perception of magnitude variations. All developed techniques are domain-agnostic and are practical alternatives for visualization designers facing large value ranges in time-series data.

The second topic explores visual model validation, a process by which users assess the fit and plausibility of statistical or machine learning models through visual inspection. While visual estimation (i.e., visual model building) has received considerable attention, fewer studies address the perception and judgment of already computed model results. This dissertation investigates the cognitive and perceptual processes involved in validating visualized linear regression models. Through experimental human-subject studies, the accuracy of visual validation and estimation is compared, and key factors are identified that influence users' ability to reliably validate model results. These factors include perceptual biases, user strategies, as well as data and design features. The findings contribute to a better understanding of visual validation processes and provide useful insights for machine learning applications and the design of visual analytics systems.

Together, this dissertation advances the theoretical and practical foundations of visual data analysis by introducing novel techniques for encoding large value ranges and empirically evaluating human assessment of model outputs. These contributions enhance the communication of complex data in high-impact domains, such as pandemic monitoring or socioeconomic analysis. Furthermore, a deeper understanding of visual model validation enables practitioners to better interpret the uncertainty of model predictions, e.g., in medical diagnostics. The findings presented provide a foundation for future research in visualization design, human-centered AI, and explainable analytics.

# *Acknowledgements*

This dissertation marks a milestone in my academic journey. Even though a PhD had been a big dream of mine ever since, I couldn't believe it would actually happen for a long time. But I really wanted to prove to myself that I was capable of doing it. I am sincerely grateful for all the inspiring people who have accompanied me on this journey.

The biggest thanks go to my supervisor, *Professor Tatiana von Landesberger*, who awoke my passion for visualization and gave me the opportunity to do my doctorate. Tatiana, thank you for supporting and guiding me through all the ups and downs of the PhD. You have always given me the freedom to pursue my own ideas, while at the same time providing competent advice whenever needed.

Furthermore, I thank *Professor Manuela Waldner* and *Professor Natalia Andrienko* for accepting to be examiner of my dissertation, as well as *Sibylle Schroll* and *Alexander Apke* for completing my doctoral committee.

I would like to express my appreciation to all my colleagues in the VisVA-Group, who have created a positive atmosphere in which I have enjoyed working every day. I particularly thank *Laura Pelchmann* and *Tom Baumgartl* for the inspiring conversations and refreshing laughs outside of work and *Max Sondag* for the helpful discussions on all the research questions.

A special thanks goes to the collaborators of my research work, who not only gave me constructive feedback on my research throughout, but also listened to every other of my academic questions: *Rita Borgo* for making sure I did not feel lost on my first conference trip, *Remco Chang* for driving me to think outside the box, and *Michael Gleicher* for having an eye on the important details.

I have been fortunate to meet many exciting people at conferences and workshops. Thank you for always having an interesting thought on my research and making me feel like a part of the visualization community: *Hans-Jörg Schulz*, *Natalia and Gennady Andrienko*, *Silvia Miksch*, and *Kresimir Matkovi*, to name just a few.

Moreover, I thank *my parents*, without whom I would not be writing these words right now. Thank you for raising me to the person I am today and motivating me to be able to achieve everything that I really want. Mom, your infinite love will always be in my heart. Dad, thanks for always being there for me and supporting me unconditionally.

Finally, I would like to deeply thank my wife *Justine*, who is always by my side, supports and motivates me, and makes me the happiest person in the world. You are my home, where I can always let go even in stressful phases.

# Contents

# List of Abbreviations and Symbols

| | |
|---|---|
| **VA** | **V**isual **A**nalytics |

*Visual Communication of Large Value Ranges*

| | |
|---|---|
| **LVR** | **L**arge **V**alue **R**anges |
| **OMC** | **O**rder of **M**agnitude **C**olors (color scale) |
| **OMC**$_{sl}$ | **OMC S**moothed **L**ightness |
| **RGB** | **R**ed **G**reen **B**lue (color space) |
| **HSV** | **H**ue, **S**aturation, **V**alue |
| **OMH** | **O**rder of **M**agnitude **H**orizon graph |
| **OML** | **O**rder of **M**agnitude **L**ine chart |
| **SSB** | **S**cale-**S**tack **B**ar chart |
| **Log** | **Log**arithmic (line chart) |
| **Lin** | **Lin**ear (line chart) |
| **HSLC** | **H**eight-**S**tack **L**ine **C**hart |
| **OMLr** | **OML** without color |
| **OMLs** | **OML S**uperimposed |
| **WS** | **W**arming **S**tripes |
| $\bar{x}$ | Mean value |
| $\tilde{x}$ | Median value |

*Visual Validation of Regression Models*

| | |
|---|---|
| **val** | **Val**idation |
| **est** | **Est**imation |
| **OLS** | **O**rdinary **L**east **S**quares regression |
| **ODR** | **O**rthogonal **D**istance **R**egression |
| **dev** | **Dev**iation |
| **CI** | **C**onfidence **I**nterval |
| **KS** | **K**olmogorov-**S**mirnov test |
| **men** | Outliers **men**tioned |
| **non** | Outliers **non**-mentioned |
| **AI** | **A**rtificial **I**ntelligence |

# Chapter 1

# Introduction

## 1.1 Background and Motivation

This dissertation investigates two important aspects of visual data analysis: the effective visual communication of large value ranges and the visual validation of regression models. Both topics are situated within the broader context of standard data analysis workflows, where visualization plays a central role in data understanding and interpretation.

Visual data analysis has emerged as a powerful approach to derive meaningful insights from complex and voluminous data sets. This analytical approach transforms raw data into visual representations to facilitate human understanding, enabling users to identify patterns, trends, anomalies, and relationships that may be difficult to detect through numerical inspection alone. Visual data analysis supports a wide range of use cases — from exploratory research to decision-making — and builds the intersection between data science, information visualization, and human–computer interaction. It amplifies human cognitive capabilities through visual representations, allowing users to process large amounts of data through perception rather than abstract reasoning alone [35, 129, 184, 193].

The diversity of visual data analysis techniques reflects its wide applicability. It spans from static visualization designs to interactive visual systems capable of handling high-dimensional, multivariate data. These techniques must often accommodate extreme value differences or computational model results. Consequently, visualization must not only convey data in an effective way but also support trust and comprehension in the face of complexity [41, 56, 66].

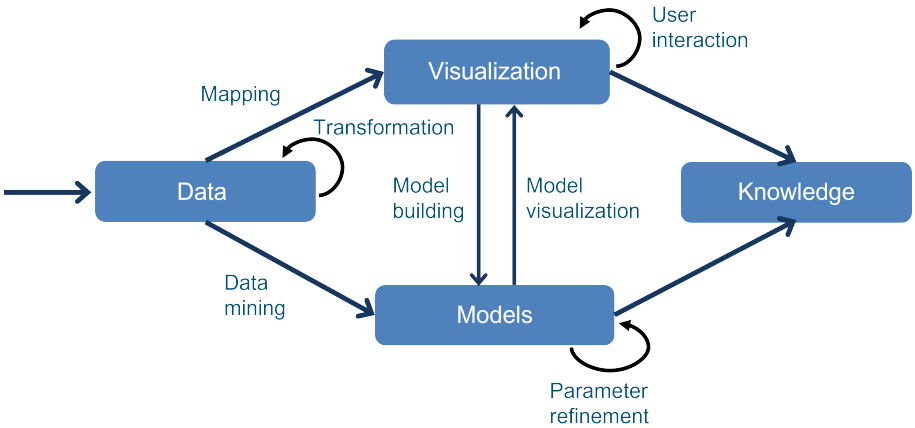



FIGURE 1.1: The visual analytics pipeline by Keim et al. [107].

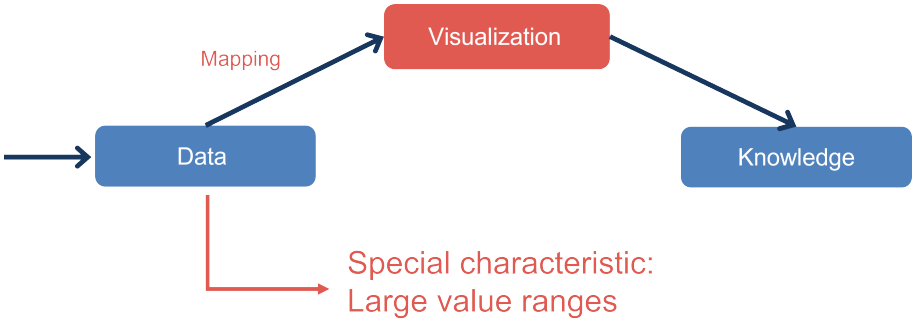



FIGURE 1.2: The parts of Keim et al.'s visual analytics pipeline relevant for the *visual communication of large value ranges*.

To guide this multifaceted process, Keim et al. [107] introduced their visual analytics (VA) pipeline, a conceptual model that outlines the key stages of visual analysis (Figure 1.1):

- *Data transformation*: Raw data is preprocessed, cleaned, and structured in a way that facilitates the subsequent analyses.
- *(Interactive) Visualization and visual mapping*: The processed data is encoded into visual forms based on chosen visual variables. The visual representations may be presented through user interfaces that support various interaction techniques to adjust the data views.
- *Model building and visualization*: Statistical models are calculated on the transformed data. The model results are visualized along with the data, allowing for their validation. Model parameter can be refined either interactively or computationally.
- *Knowledge generation*: Through iterative exploration and interpretation, users derive insights, make decisions, or generate new hypotheses.

This pipeline not only delineates the stages of visual data analysis but also emphasizes the recursive nature of the process. Rather than a strictly linear sequence, the pipeline allows feedback loops in which users refine earlier steps based on insights gained during interaction, underscoring the importance of human involvement at each stage. It also serves as a useful framework for positioning specific challenges and design strategies within the broader scope of visual analytics.

The research conducted in this dissertation covers different parts of the visual analytics pipeline to explore two key challenges in visual data analysis. Both topics highlight essential, yet often under-addressed, aspects of designing effective visualizations — how to accurately and accessibly present data with high numerical variation, and how to support users in understanding and assessing the validity of computational models through visualization.

The first topic — the **visual communication of large value ranges** — focuses on novel visual designs and improved color coding that enhances both the readability of mantissa values as well as the perception of magnitude variations. This topic relates to the visual mapping process of the VA pipeline, in which problem specific encodings are developed for an accurate interpretability of the underlying data. The defining characteristic of the investigations is that the visualized data contain large value ranges (see Figure 1.2).

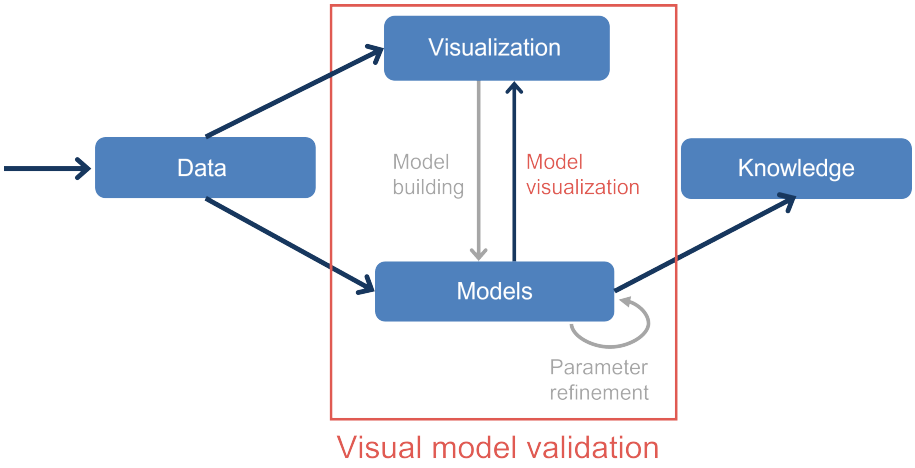



FIGURE 1.3: The parts of Keim et al.'s visual analytics pipeline relevant for *visual model validation*.

In the second topic — the **visual model validation** —, the interplay between computationally derived statistical models and human perception is examined. By juxtaposing model results with raw data, analysts can iteratively refine both their models and their understanding of the underlying processes. Reliable visual validation of the model results is essential for this process. Hence, this work studies people's ability to visually assess the quality of a model and the factors that influence this, such as perceptual bias and visual design.

As shown in Figure 1.3, visual model validation is only implicitly incorporated into standard visual analytics frameworks. Therefore, in this thesis, the role of visual model validation in visual analytics processes is explicitly described and integrated into a novel VA pipeline.

Altogether, the two fields studied in this dissertation advance both the design space of large value ranges visualization and the transparency of analytic modeling. Thereby, they strengthen individual parts of the visual analytics process.

## 1.2 Visual Communication of Large Value Ranges

### 1.2.1 Background, Motivation, and Related Work

Data sets with large value ranges (LVR) are characterized by values that encompass multiple exponents (i.e., orders of magnitude) as represented using the *scientific notation* [17, 23, 93, 94]. In this notation, each value $v$ is divided into its mantissa $m \in \mathbb{R} \setminus \{0\}$ and exponent $e \in \mathbb{Z}$:

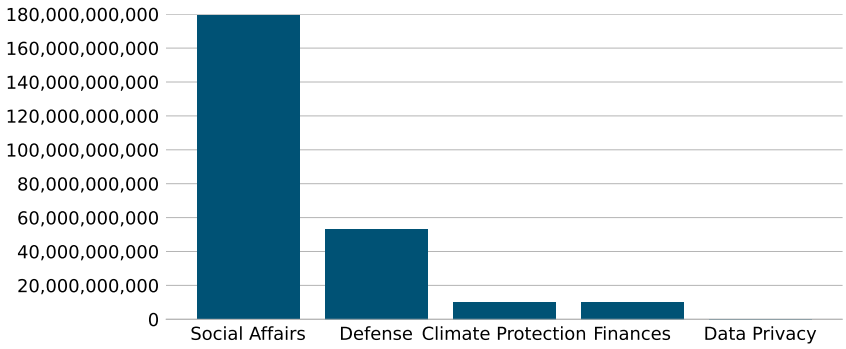
$$v = m \cdot 10^e \tag{1.1}$$

In this dissertation, the most common base 10 is used, but other bases (e.g., base 2 for floating-point numbers) are also possible. The *scientific normalized notation* is employed, in which $e$ is chosen so that $|m| \in [1, 10)$ [18].

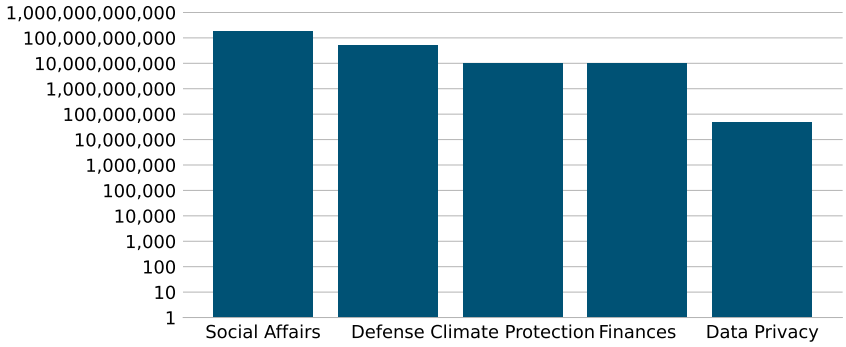
Large value ranges have become an integral part of various scientific fields (e.g., cloud measurements in meteorology [162] or physical measurements such as the electromagnetic spectrum [204]) as well as several domains of daily life (e.g., pandemic tracing [121] or cryptocurrency or stock prices for financial investments [133]). Therefore, an effective static visualization of large value ranges is important for both scientists and the public to provide an informative overview, both for print and digital media.

Standard visualization techniques, such as linear or logarithmic scaled charts, work well for small value ranges, but prove inadequately for large value ranges. The example in Figure 1.4 illustrates that a linear scale makes it extremely difficult to read and compare values in lower orders of magnitudes. Values below a certain threshold are even rendered at sub-pixel sizes, resulting in a loss of information (see Figure 1.4a). The utilization of a logarithmic scale is a viable solution to this issue, as it ensures that all values are rendered distinctly (see Figure 1.4b). Nevertheless, it introduces significant interpretative challenges. Logarithmic representations are not widely understood by the general public [47, 125, 155] and have been shown to cause misinterpretations even among scientific audiences [111]. Additionally, logarithmic scales complicate fundamental analytical tasks, such as accurately estimating ratios between values of similar or adjacent magnitudes, thereby reducing their effectiveness for detailed quantitative comparison [93].

Previous research has introduced alternative visualization techniques to overcome the limitations of linear and logarithmic scales. These approaches commonly decompose values into their mantissa and exponent, which are then visualized using distinct visual encodings or novel scales. Hlawatsch et al. [93], Borgo et al. [23], and Höhn et al. [94] each introduced adaptations of the traditional bar chart to accommodate data with large value ranges. These approaches employed distinct visual variables to represent univariate data values. Their studies demonstrate that incorporating the unique characteristics of large value ranges into visualization design enhances performance on tasks such as value retrieval, sorting, ratio and difference estimation, and trend detection. However, their investigations have been limited to nominal data and variations of the traditional bar chart.




(A) Linear scaling.




(B) Logarithmic scaling.

FIGURE 1.4: Linear bar chart (A) and logarithmic bar chart (B) visualizing a sample of the German government's budget allocations (in €) [60], illustrating order of magnitude differences.

### 1.2.2 Research Questions

Large value ranges are not only found in nominal, one-dimensional data. This dissertation proposes novel visualization designs aimed at representing large value ranges in multivariate data types. Specifically, the included works address different variants of time-dependent data, offering novel perspectives on its visual communication. The difficulty with time-dependent data is to depict the individual data points with clarity while preserving the continuous nature of temporal relationships. Accordingly, this thesis answers the following research questions:

- **RQ1.1**: Do visualization designs specially developed for large value ranges in time-dependent data improve the readability and comparability of such data?
- **RQ1.2**: How does the visual mapping to color contribute to the readability of magnitude variations in large value range visualizations?

Based on the prior research, it can be assumed that considering the characteristics of large value ranges in the development of visualizations for time-dependent data results in enhanced effectiveness compared to respective standard visual designs. Color is widely recognized as one of the most effective visual variables in data visualization, as it significantly enhances both the readability and perceptual clarity of visual information [19,193]. It has the ability to convey categorical as well as sequential distinctions and to support pre-attentive processing [129]. That makes it an effective element in communicating differences both at a high-level — i.e., variations in the orders of magnitude — and at a low-level — i.e., distinctions of mantissa values within individual magnitudes.

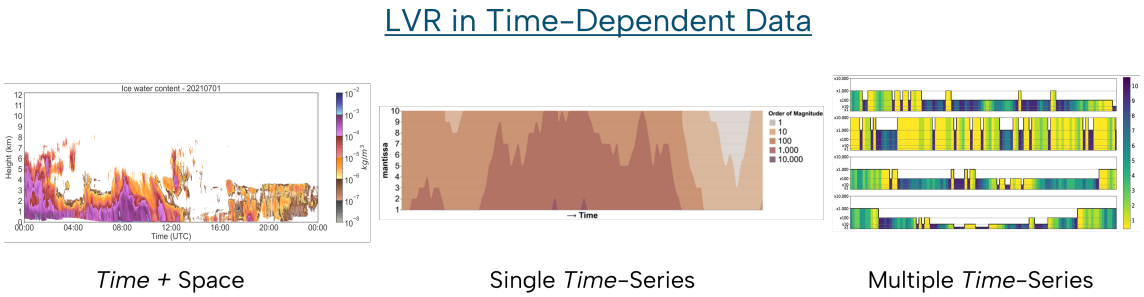
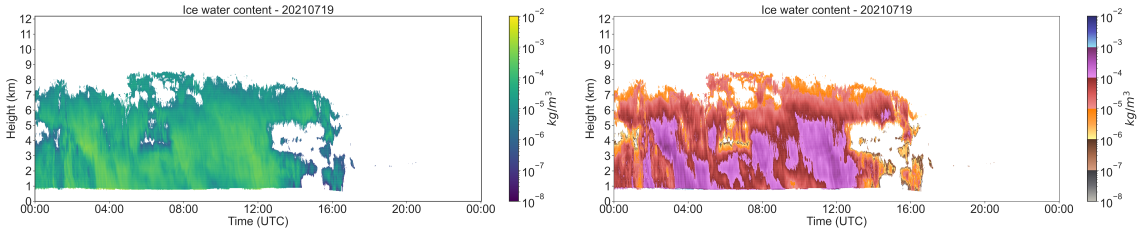



FIGURE 1.5: Overview of the research on visual communication of large value ranges included in this dissertation.




(A) The commonly used color scale *viridis*.

(B) The new *order of magnitude color scale*.

FIGURE 1.6: Exemplary comparison of the new OMC **color scale for large value ranges** with the previously used standard viridis color scale [29]. The color scales are applied to meteorological cloud data. It is noticeable that with the new scheme the differences in magnitude are significantly easier to perceive.

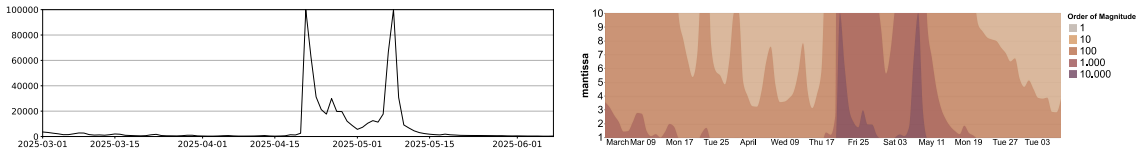
### 1.2.3 Contributions

The predominant focus of earlier studies on visually encoding large value ranges has been on nominal data representations. In contrast, the work presented in this dissertation extends this research frontier by exploring visualization designs for time-dependent data (see Figure 1.5) — an area that introduces additional complexity due to temporal variation and continuity [5].

In this work, separate representations of mantissa and exponent are used. Previous approaches are significantly extended by introducing novel visual encoding channels to meet the unique challenges of time-series data. In summary, this dissertation advances the state-of-the-art through the following contributions:

- A novel "nested" **color scheme for large value ranges** is proposed (see Figure 1.6). It leverages the numerical decomposition of values into mantissa and exponent components to highlight differences between magnitude. The approach is applied to meteorological spatio-temporal data and compared with three state-of-the-art alternatives in an empirical user study. The results demonstrate that the new color scheme significantly increases accuracy, response time and confidence in interpretation tasks, while maintaining comparable effectiveness to existing schemes in discrimination tasks. The corresponding color coding concept is implemented in the open access Python library omccolors. (Section 2.1) [29]

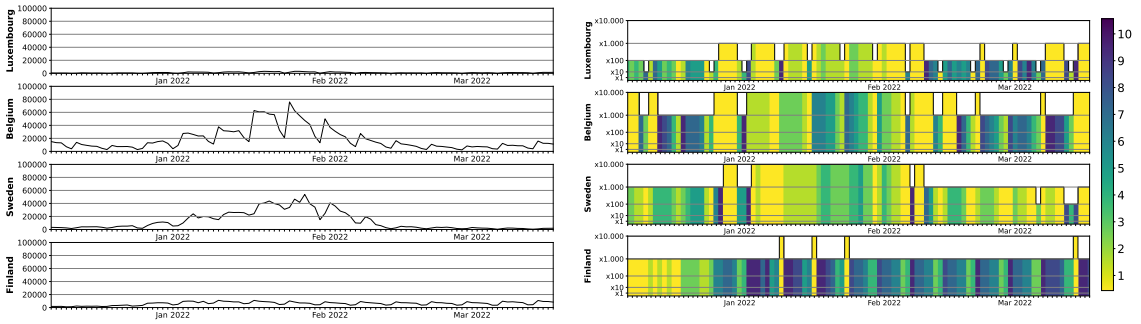
- **Two new visualization techniques for individual time-series data containing large value ranges** are presented that adapt and extend the ideas of standard visualization designs (see Figure 1.7). Furthermore, a **novel scale** is introduced that offers the benefits of both the logarithmic scale's visibility of all values and the linear




(A) Standard linear line chart.

(B) The novel *order of magnitude horizon graph*.

FIGURE 1.7: The figure compares the novel order of magnitude horizon graph and the common linear line chart for **large value ranges in individual time-series** [25]. The data represent page views for the article "Papstwahl" on de.wikipedia.org. The proposed visualization technique facilitates value identification and comparison even within lower orders of magnitude, which are often less discernible in standard representations.




(A) Standard linear line chart.

(B) The novel *height-stack line chart*.

FIGURE 1.8: Application example of the new visualization design for **large value ranges in multiple time-series** compared to the standard linear line chart [26]. Both designs show the number of daily new COVID-19 cases compared for different countries. Due to the reduced available space, a detailed comparison of time-series is difficult using standard visualization techniques.

scale's intuitive readability. An empirical user study was conducted to evaluate the effectiveness of these designs in comparison to state-of-the-art visualization methods. The results show that the novel order of magnitude horizon graph consistently outperforms or matches existing techniques in tasks involving identification, discrimination, and estimation. (Section 2.2) [25]

- The challenging problem of **visualizing large value ranges across multiple time-series** is addressed (see Figure 1.8). To this end, a **design space for large value ranges** in time-series data and their composite visualizations is proposed. Seven distinct visualization designs are evaluated through crowdsourced user studies: three representative state-of-the-art techniques, three extensions of existing methods, and one novel design. For the task of minimum value identification, the novel height-stack line chart outperforms all other designs. In tasks in which the maximum value serves as an effective proxy for the correct answer, the traditional linear line graph demonstrates performance comparable to all other evaluated designs. Contrary to common assumptions, the results suggest that increasing the number of time-series does not significantly impair accuracy in estimation, discrimination, or identification tasks. (Section 2.3) [26]

| *Reference* | *Data Type* | *Mark Type* | *Range for Exponent* | *Channel for Exponent* | *Channel for Mantissa* | *Task Evaluation* |
|---|---|---|---|---|---|---|
| SSB [93] | nominal | bar | 5 ($10^0$ – $10^4$) | Row | Y-Position | Value, Difference, Ratio |
| OMM [23] | nominal | bar | 5 ($10^1$ – $10^5$) | Y-Position and Color Hue | Y-Position and Color Hue | Value, Sort, Ratio, Trend |
| WSB [94] | nominal | bar | 5 ($10^0$ – $10^4$) | Width and Color Intensity | Y-Position | Value, Sort, Ratio, Trend |
| **OMC [29]** | time-series | area | 7 ($10^{-8}$ – $10^{-2}$) | Color Hue | Color Intensity | Value, Sort |
| **OMH [25]** | time-series | area | 5 ($10^0$ – $10^4$) | Y-Position and Color Hue | Y-Position and Color Intensity | Value, Sort, Difference, Trend |
| **OML [25]** | time-series | area | 5 ($10^0$ – $10^4$) | Color Hue | Y-Position | Value, Sort, Difference, Trend |
| **HSLC [26]** | time-series | area | 5 ($10^0$ – $10^4$) | Y-Position | Color Intensity | Value, Sort, Difference, Ratio |

TABLE 1.1: Overview of large value ranges visualization techniques proposed in the literature (adapted and extended from Batziakoudi et al. [18]).

Importantly, all findings and developed techniques are domain-agnostic and are applicable to any context involving time-series data with large value ranges, holding practical relevance for visualization designers. Possible application areas include scientific fields such as geophysics (e.g., energy release in earthquakes [104]) or astrophysics (e.g., properties of interplanetary solar flares [103]) but also fields of public interest like socioeconomic trends (e.g., income inequalities in the US [143]) or transportation (e.g., individual mobility patterns [82]).

The novelty and significance of the contributions of this thesis in evolving the current understanding and development of techniques for the visual communication of large value ranges is showcased by Batziakoudi et al. [18]. Their overview of existing research on the visualization of large value ranges highlights that more than half of the contributions in this research domain originate from the papers included in this thesis (see Table 1.1).

The chapters and contributions in this dissertation on the visual communication of large value ranges are based on the following research publications:

- **D. Braun**, K. Ebell, V. Schemann, L. Pelchmann, S. Crewell, R. Borgo, and T. von Landesberger. Color coding of large value ranges applied to meteorological data. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 125–129, 2022. doi: 10.1109/VIS54862.2022.00034
- **D. Braun**, R. Borgo, M. Sondag, and T. von Landesberger. Reclaiming the horizon: Novel visualization designs for time-series data with large value ranges. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1161–1171, 2024. doi: 10.1109/TVCG.2023.3326576
- **D. Braun**, R. Borgo, M. Sondag, and T. von Landesberger. Design and Evaluation of Visualizations for Large Value Ranges in Multiple Time-Series. *Information Visualization*, 2025. doi: 10.1177/14738716251349501

## 1.3 Visual Validation of Regression Models

### 1.3.1 Background, Motivation, and Related Work

Model validation is the process of assessing whether a model accurately represents the underlying data. This process can be conducted through computational methods or visual analysis. In computational validation, statistical metrics are computed to quantify model performance and determine whether the results meet predefined criteria. In contrast, visual validation involves the graphical representation of the model result alongside the data, enabling humans to assess the model's fit to the data.

Thus, visual model validation can be defined as the evaluation of a model's behavior, structure, or outputs through visual representations, in order to assess its accuracy, coherence, and alignment with real-world or theoretical expectations. This approach leverages the human capacity for pattern recognition and anomaly detection, allowing experts and stakeholders to identify inconsistencies, errors, or areas of concern.

Visual model validation is a foundational component of exploratory data analysis [49, 96]. Although traditional statistical metrics offer quantitative measures of model performance, they often obscure nuanced characteristics in complex models and data sets. Indeed, data sets with markedly different structures can yield nearly identical statistical summaries [21, 122, 181]. In such cases, visual validation becomes indispensable for revealing subtle patterns and anomalies that would otherwise go unnoticed.

Due to the inherent complexity of many statistical models — such as regression analyses — visual inspection is frequently used to evaluate their correctness and reliability [40, 41]. For instance, machine learning model outputs generated on training data are commonly evaluated through visual comparison with results from a validation set [197], or global patterns may be contrasted with regional data to identify local deviations [182]. This practice is crucial in high-stakes domains, including pandemic monitoring [55] and meteorological forecasting [102], where model predictions can have significant real-world consequences. Moreover, the general public engages in visual validation when interpreting charts and forecasts in news media or on television (e.g., weather trends in Figure 1.9).

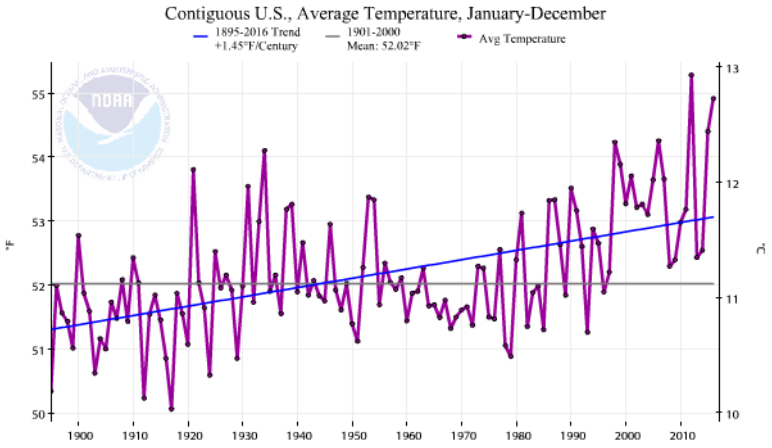



FIGURE 1.9: Annual average temperature for the contiguous United States from 1895-2016 (shown in [54]). The chart serves as an example for a model result/ trend to be validated by a viewer in a news media setting.

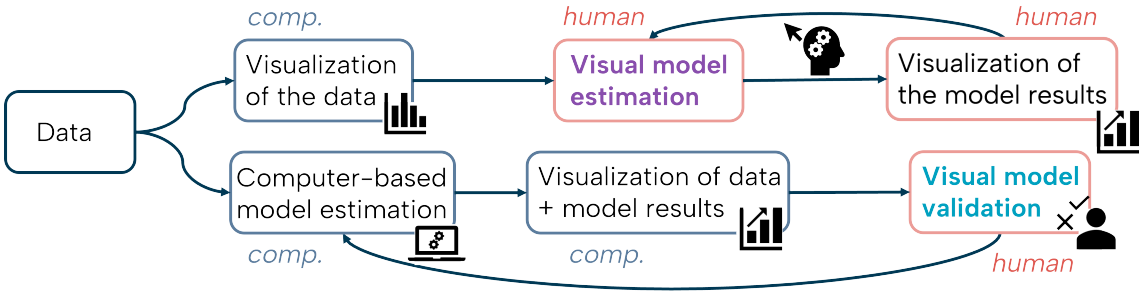



FIGURE 1.10: Schematization of the *visual model validation* and *estimation* processes. Note the change in the distribution of human and computer influence on model generation.

Visual model validation represents a subsequent stage in the visual analysis process, wherein previously generated model results are assessed by humans. The corresponding model generation can also be performed either by computational algorithms or manually by human analysts. In the case of *visual model estimation*, the user takes an active role in generating the model based on visual inspection of the data. This can occur in various forms. For example, a data scientist might first visually examine a plot to decide which type of model to apply. A more interactive variant of visual estimation involves users directly manipulating model parameters on a display (e.g., as part of a visual analytics system) until the model appears to fit the data [46, 52, 80, 95].

Figure 1.10 illustrates the differences between these two processes. A significant difference is the influence of computers and humans on model generation: The computer calculates the model parameters before the visual validation, while the user interprets the visualized data and mentally constructs a plausible model during visual estimation.

Importantly, visual estimation inherently involves a validation element. Each manual adjustment to a model is accompanied by a judgment about whether the fit has improved or whether further refinement is needed. However, this should be distinguished from visual model validation as an independent task. In visual validation, users compare a given (often computer-generated) model against their own internal understanding of the data. Here, the focus is not on creating the model, but rather on evaluating its plausibility or accuracy based on prior knowledge or observed patterns.

The topic of visual model validation has received limited attention within the research community in the past. Only few works have dealt with the human ability to visually validate model results. Majumder et al. [119] investigated visual validation in the context of statistical inference for linear models. Similarly, Correll et al. [50] examined how users visually validate data distributions across various visualization types. In practice, opportunities for visual model validation are most often embedded within interactive visual analytics and machine learning systems [22, 40, 42, 43, 130].

The majority of studies on the perception of statistical models were focused on understanding visual estimation. The average value [80, 95, 199], linear regression models [46, 52], and the correlation of two data dimensions [85, 106, 152, 181, 196, 198], are among the models for which humans' estimation ability has already been investigated. Most of these works studied perceptual biases, data properties, and visual features that influence the visual model estimation process.

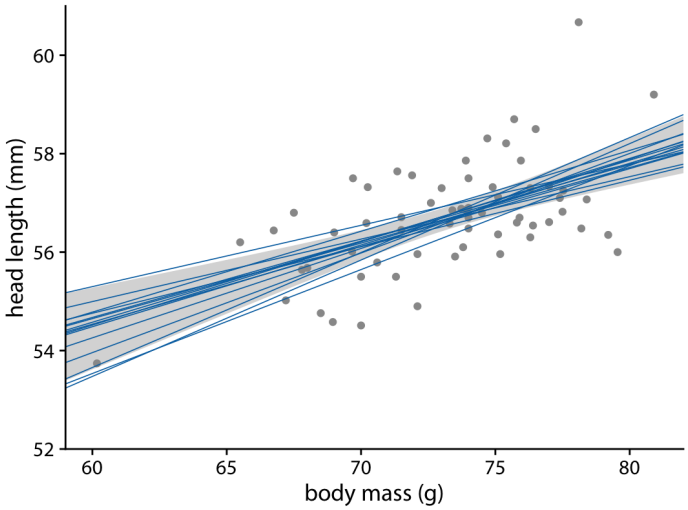



FIGURE 1.11: Head length versus body mass for male blue jays. The gray band represents a 95% confidence level. The blue lines represent statistically equally likely alternative fits randomly drawn from the posterior distribution (shown in [195]).

### 1.3.2 Research Questions

This dissertation examines the human ability to visually validate statistical models. While visual model estimation has been widely researched, this work is the first to systematically examine perceptual biases, data characteristics and visual attributes that affect visual model validation. To establish a controlled foundation for the investigations, the presented studies focus on linear regression models, as these are commonly used in both academic and applied settings and are conceptually easily accessible also for non-experts.

A *linear regression model* estimates the linear relationship between two data dimensions — the dependent variable $y \in \mathbb{R}$ and the independent variable $x \in \mathbb{R}$:

$$y = a \cdot x + b + \epsilon \tag{1.2}$$

It estimates the regression coefficients (i.e., the slope value $a \in \mathbb{R}$ and intercept $b \in \mathbb{R}$ for linear regression) such that the error term $\epsilon \in \mathbb{R}$ is minimized. Linear regression models can be used to identify trends and central tendencies.

However, these trend estimates are subject to uncertainty in their coefficients. Statistically, this implies that multiple linear models may be consistent with the observed data, rather than a single definitive fit (see Figure 1.11). Consequently, it becomes important to examine the extent to which individuals can discern between statistically valid and invalid trend lines.

The standard ordinary least squares (OLS) model is used for the regression estimations. This model minimizes the sum of the squared residuals, i.e., the vertical distances between the observed dependent variable (values of the variable being observed) in the input data set and the output of the model with respect to the independent variable (see Figure 1.12).

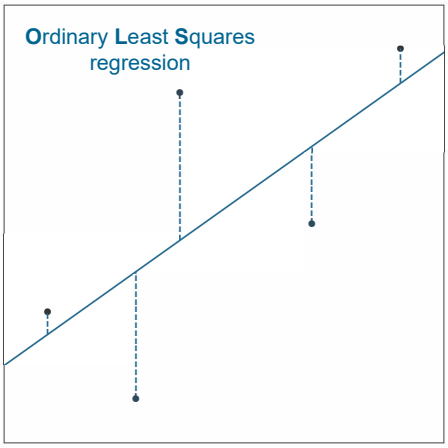



FIGURE 1.12: Example for OLS regression modeling. The error lines illustrate that the model minimizes the (squared) vertical distances between the data points and the regression line.

By restricting the conducted studies to this well-defined model class, this dissertation systematically answers the following research questions:

- **RQ2.1**: Are individuals able to perform visual validation consistently and without bias for (linear) regression models?
- **RQ2.2**: How does performance in visual model validation relate to the accuracy of visual model estimation (in scatterplots)?

These questions are addressed through human-subject experiments in which participants are shown scatterplots overlaid with trend lines representing regression model outputs. The parameters of these models (i.e., parameters $a$ and $b$ in Equation 1.2) are systematically varied, such that only a subset of the displayed trend lines accurately reflect the underlying data. Participants are asked to decide whether they accept or reject each line as the correct trend of the data. This methodology allows us to assess the degree of deviation from the true model parameters that participants tolerate before identifying a model as incorrect.

Due to its task structure, which involves accepting or rejecting predefined model results, visual validation should comparatively be more intuitive for humans than visual estimation, which necessitates more in-depth decisions regarding model adaptations. Moreover, as discussed in Subsection 1.3.1, visual estimation typically includes an implicit visual validation step. This suggests that visual validation is cognitively less demanding than visual estimation and should yield more accurate results.

### 1.3.3 Contributions

With the work on visual model validation, this thesis makes a substantial contribution to the understanding of how humans visually process statistical information. The strengths and limitations of relying on visual assessment of regression models are examined through a comprehensive conceptual and empirical investigation of visual model validation. Moreover, this work clarifies the visual analysis and modeling process methodologically by distinguishing and comparing the processes of visual validation and estimation across a progression of regression models.

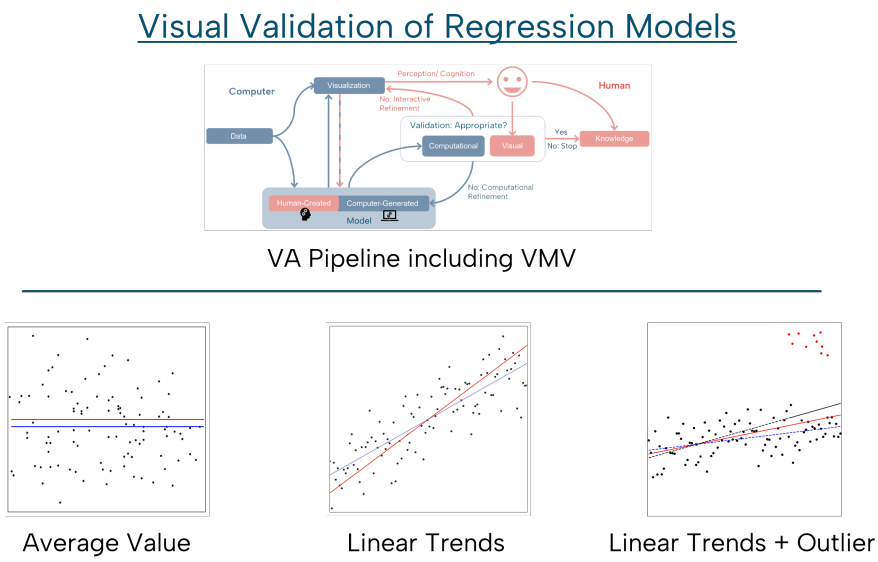



FIGURE 1.13: Overview of the research on visual validation of regression models included in this dissertation.

In particular, this dissertation includes the following contributions (see Figure 1.13):

- As seen in Figure 1.3, visual model validation is no integral part of the classic visual analytics pipeline. Advances in the field of visual analytics are made by explaining the role of visual model validation and estimation in VA processes. A **novel VA pipeline** is proposed that seamlessly integrates validation and estimation loops into the conceptual structure of existing frameworks. Through detailed exposition and multiple case studies, it is demonstrated how the new pipeline applies across various modeling tasks and VA systems. (Section 3.1)

- The empirical investigation on the visual validation of regression models is started with a human-subject study on the mean (i.e., linear regression to a constant) as a common **model of central tendency**. The study involved two distinct participant groups — crowdsourced workers and volunteer participants. The accuracy of visual validation and estimation and their performance against statistical results are compared. The findings reveal that the accuracy of models deemed valid by participants was systematically lower than the accuracy of those they estimated themselves. Importantly, it is shown that participant responses in both estimation and validation tasks were largely unbiased. Figure 1.14 illustrates the results of the study. (Section 3.2) [30]

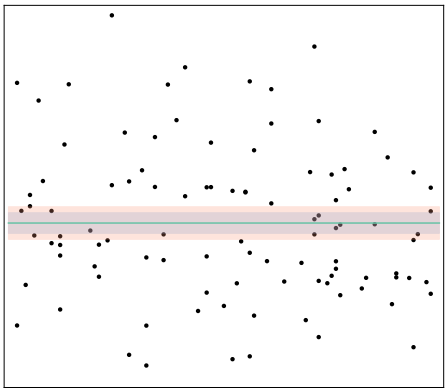


FIGURE 1.14: Summary of the results of the empirical investigation on the **visual validation of the average value** presented in an example stimulus [30]. The figure shows the true average value (green) together with participants' average estimation range (blue) and the range of lines with an acceptance rate of 50% or higher for validation (orange). It stands out, that the visual estimation results are slightly more accurate than the validation results.

- Subsequently, the model complexity is increased and the efficacy and limitations of **visual model validation for linear regression** in scatterplots is examined. Through two experiments, individuals' abilities to validate (and estimate) linear regression models is investigated and the influence of common visualization design elements on validation performance is evaluated. The first experiment found that participants were more accurate when visually estimating regression slopes than when validating pre-defined slopes. Interestingly, participants consistently exhibited a bias toward slopes that were steeper than the ground truth (see Figure 1.15). The second experiment showed that widely used visualization aids — such as error lines, bounding boxes, and confidence intervals — do not enhance visual validation accuracy. (Section 3.3) [27]

- The **visual model validation in the presence of outliers** is studied, a common but complex scenario in regression analysis that requires implicit decisions about how to treat those outliers. Two quantitative human-subject studies investigate the mental models individuals apply when visually validating linear regression and how their decisions are shaped by the presence of outliers and other data characteristics. The first study compares visual model validation to visual model estimation. The second study investigates validation behavior and its influencing factors in more detail. The results demonstrate that participants are less consistent and more sensitive to data variability during validation than estimation, particularly in how they handle outliers (see Figure 1.16). It is also shown that individuals tend to adopt a personal strategy — either consistently including or excluding outliers — regardless of whether outliers are explicitly discussed. These strategies are influenced primarily by the directional congruence of outliers. (Section 3.4) [28]

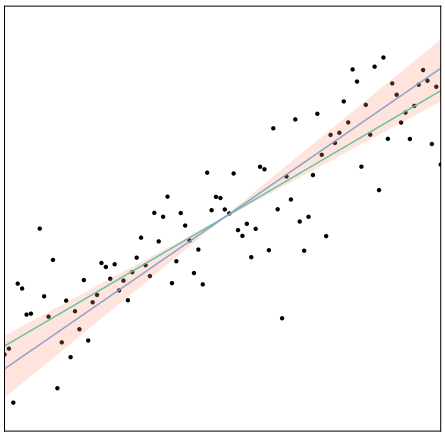


FIGURE 1.15: Illustrative example for one of the main results of the first study on the **visual validation of the linear trends** [27]. The figure shows the true trend line (green) together with participants' average trend estimation (blue) and the range of lines with an acceptance rate of 50% or higher for validation (orange). In both tasks, participants most likely accepted and drew lines that were too steep.

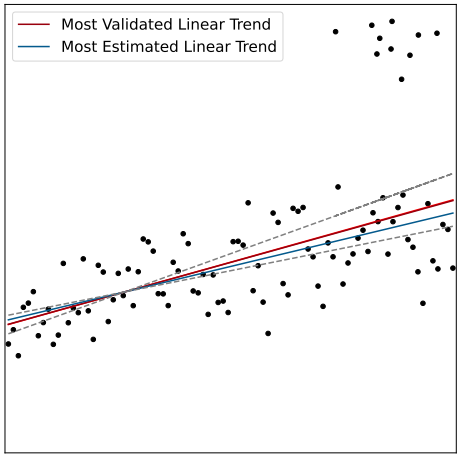



FIGURE 1.16: Summarizing the results for the **visual validation and estimation of liner trends with outliers** [28]. The figure displays the most accepted and estimated trend lines compared to the OLS regression models including and excluding the outliers (dashed lines). It shows that outliers are more likely to be considered in visual validation than in estimation.

The work in this dissertation lays the groundwork for understanding visual model validation, which has implications for a range of practical applications in which human judgment interacts with statistical modeling. For instance, insights into biases and strategies in human visual validation can guide the creation of adaptive visualization interfaces that offer corrective feedback or context-aware design adjustments. The proposed VA pipeline and empirical findings also support the development of decision support systems in domains such as healthcare [114] or finance [14], where users must validate predictive models visually to make high-stakes decisions. Moreover, understanding how people visually interpret regression models can inform the design of visualization aids that better align with human intuition. Overall, this work informs the building of VA systems that not only present data visually but also account for — and enhance — the user's ability to reason about model correctness.

The majority of the contributions on the visual validation of regression models are based on the following research publications and papers in preparation for submission:

- **D. Braun**, A. Suh, R. Chang, M. Gleicher, and T. von Landesberger. Visual validation versus visual estimation: A study on the average value in scatterplots. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 181–185, 2023.
  doi: 10.1109/VIS54172.2023.00045
- **D. Braun**, R. Chang, M. Gleicher, and T. von Landesberger. Beware of validation by eye: Visual validation of linear trends in scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):787–797, 2025.
  doi: 10.1109/TVCG.2024.3456305
- **D. Braun**, D. Eberle, R. Chang, M. Gleicher, and T. von Landesberger. Deciding Through Noise: Visual Validation of Linear Trends in Scatterplots Amid Outliers. *In preparation for submission to CHI conference*, 2026.

## 1.4 Outline of the Thesis Structure

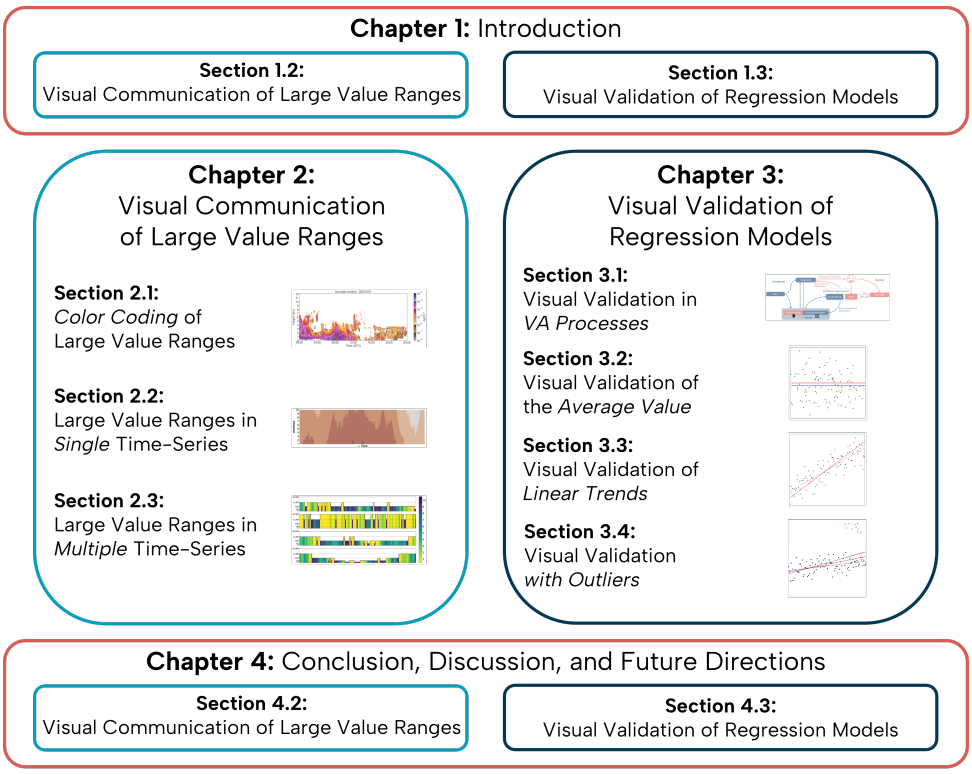



FIGURE 1.17: Outline of the thesis structure.

The remainder of this thesis is organized as follows (see Figure 1.17).

All papers included in this cumulative dissertation are preceded by their full bibliographic reference, a list of supplementary materials with its availability, and a statement of each author's contribution to the work.

**Chapter 2** contains the results of the research on the visual communication of large value ranges. The three papers presented deal with different facets of time-dependent data: Starting with a new coloring method for spatio-temporal data (Section 2.1), through the development of new visualization designs for large value ranges in individual time-series (Section 2.2), to the extension of these methods to multiple time-series (Section 2.3).

**Chapter 3** begins with an introduction of a novel VA pipeline that integrates and explains the role of visual model estimation and validation in visual analytic processes (Section 3.1). The remainder of this chapter includes three papers that focus on the visual validation of regression models in scatterplots. Each paper increases the complexity of the investigated model or data: First, humans' ability to visually validate the average value is examined (Section 3.2). Subsequently, the complexity of the model is increased and the perception of linear trends is investigated (Section 3.3). The most recent work additionally changed the data structure and tested how people consider outliers in their visual validation of linear trends (Section 3.4).

In Chapter 4, the findings of the two main topics are summarized and discussed. Moreover, limitations of the presented work are outline and directions for future research are suggested.

# Chapter 2

# Visual Communication of Large Value Ranges

## 2.1 Color Coding of Large Value Ranges Applied to Meteorological Data

The first publication on the visual communication of large value ranges is based on this author's master thesis "Farbabbildung von Large Value Ranges bei meteorologischen Wolkendaten", submitted to the University of Cologne in March 2022.

The resulting paper on the development of a nested color scheme that uses a separate representation of the mantissa and exponent of a value to encode large value ranges was published and presented at the IEEE Visualization conference:

**D. Braun**, K. Ebell, V. Schemann, L. Pelchmann, S. Crewell, R. Borgo, and T. von Landesberger. Color coding of large value ranges applied to meteorological data. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 125–129, 2022.
doi: 10.1109/VIS54862.2022.00034

The supplementary material of the paper, including the meteorological data, the RGB values of the OMC color scales, and the study results and documentation, is publicly available at OSF. In addition, the concepts examined in the paper are implemented in the open access Python library omccolors. The library automatically applies the OMC structure to standard Matplotlib color scales. The user just has to specify the base color scheme and desired value range.

I am the primary author of this publication. In this role, I was responsible for the design, implementation, data collection and analysis, as well as the writing and publication of the work. The specific contributions of myself and my co-authors to this publication are outlined below according to the Contributor Roles Taxonomy (CRediT):

**D. Braun**: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Software, Visualization, Writing – Original Draft. *K. Ebell, V. Schemann, S. Crewell*: Resources. *L. Pelchmann*: Writing – review & editing. *R. Borgo, T. von Landesberger*: Supervision, Conceptualization, Methodology, Writing – review & editing.

# Color Coding of Large Value Ranges Applied to Meteorological Data

DANIEL BRAUN[1], KERSTIN EBELL[1], VERA SCHEMANN[1], LAURA PELCHMANN[1], SUSANNE CREWELL[1], RITA BORGO[2], TATIANA VON LANDESBERGER[1]

[1]*University of Cologne*
[2]*King's College London*

**Abstract:**

This paper presents a novel color scheme designed to address the challenge of visualizing data series with large value ranges, where scale transformation provides limited support. We focus on meteorological data, where the presence of large value ranges is common. We apply our approach to meteorological scatterplots, as one of the most common plots used in this domain area. Our approach leverages the numerical representation of mantissa and exponent of the values to guide the design of novel "nested" color schemes, able to emphasize differences between magnitudes. Our user study evaluates the new designs, the state of the art color scales and representative color schemes used in the analysis of meteorological data: ColorCrafter, Viridis, and Rainbow. We assess accuracy, time and confidence in the context of discrimination (comparison) and interpretation (reading) tasks. Our proposed color scheme significantly outperforms the others in interpretation tasks, while showing comparable performances in discrimination tasks.

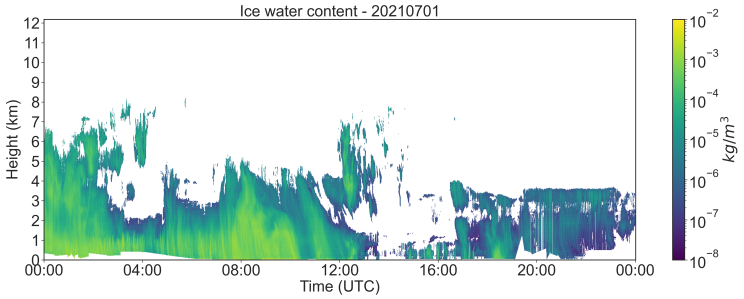
*IEEE Visualization Conference, 2022*

## 1. Introduction

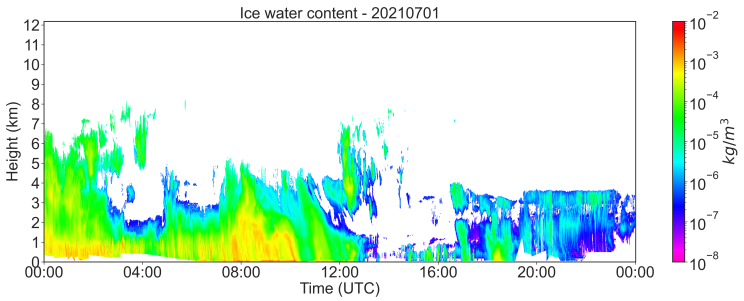
Data with large value ranges are data sets whose values contain several different exponents using the scientific notation ($v = m \cdot 10^e$) [94]. One example is meteorological data on ice water content in clouds, whose exponents vary between minus eight and minus two. The data includes three variables: time, height and ice water content. Mapping the values to position is not possible in a 2D-visualization. The current meteorological standard is to visualize the data using a scatterplot with the values encoded by a logarithmic scaled colormap [74, 97]. The meteorologists call this *time-height series*. The challenge is to encode such a large value range on a color scale and a limited number of pixels without loss of information.

The common colormaps used for meteorological scatterplots are *Viridis* (Figure 1a) [74] and *Rainbow* (Figure 1b) [97]. We present a novel color scheme to visualize data with large value ranges: the *order of magnitude colors* (OMC), which uses a separate representation of the mantissa and exponents of the data (Figure 1d) for an easier classification of the orders of magnitude. It uses one hue for each exponent and linear gradient of brightness within for magnitude. Additionally a variation of this design is shown (Figure 1e).

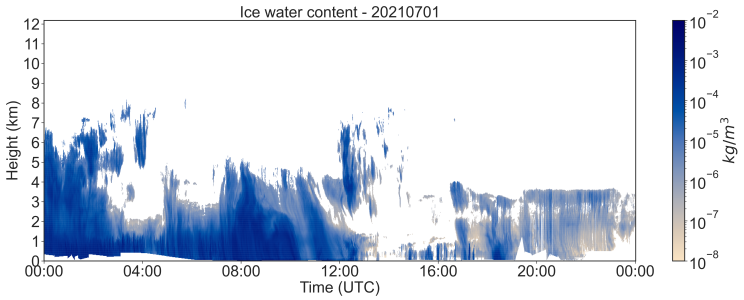
We compare our approach with the currently used color scales as well as a state-of-the-art color scheme generated via the tool *ColorCrafter* [171] (starting color: blue) (Figure 1c) in an empirical user study. The results show that our new color scheme has comparable results for comparison tasks and works significantly better for reading values than the other colormaps. We already received requests from meteorologists for the use of our design.



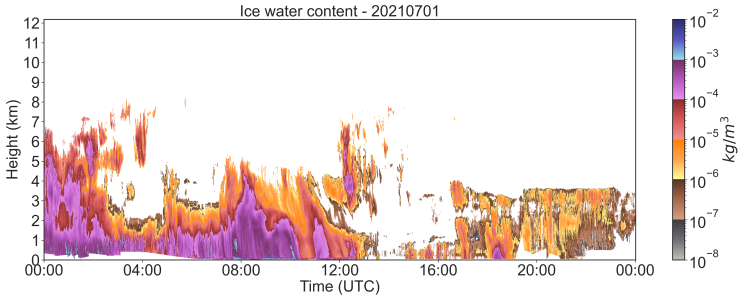

(A) Viridis



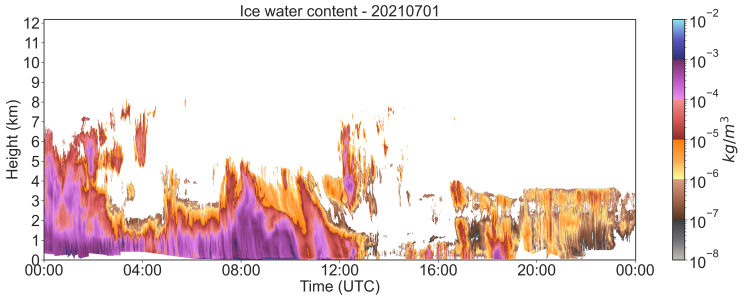

(B) Rainbow




(C) ColorCrafter [171]




(D) Order of magnitude colors (OMC)




(E) OMC smoothed lightness ($OMC_{sl}$)

FIGURE 1: Scatterplots of meteorological data for the ice water content in clouds (time height series) featuring large value ranges encoded using the color schemes Viridis, Rainbow, ColorCrafter and two versions of our *order of magnitude colors* design.

## 2. Related Work

**Color Perception** Golbiowska et al. [83] explored the perceptual differences from rainbow to sequential color scales. Their study showed that the sequential colormaps performed better in comparing values and recognizing general patterns, whereas the rainbow scale supports the reading of specific details. We also compare these two types of color schemes for the investigated tasks.

The pros and cons of the rainbow color scheme have been widely discussed in literature. Rogowitz and Treinish [153] criticize the Rainbow-colormap for the fact that the boundaries of the different colors can be perceived as boundaries in the data. This is countered by the findings of Reda and Szafir [149] and Reda et al. [148] that the more unique colors represented, the better the plot.

There are a variety of tools that can be used to generate and optimize color scales. The *ColorCrafter*-tool [171] creates algorithmic generated sequential color schemes for given parameters. These colormaps work well for the most quantitative data, but hide details when visualizing data with large value ranges. Another software is the *CCC*-tool, which allows to create completely new color schemes and to optimize given color scales [131].

A data-dependent adaptation of color schemes based on statistical properties [163, 183] would lead to incomparability of daily views. Thus, we disregard them.

**Visualization of Large Value Ranges** Most of the research work on large value ranges introduced novel visualization types for two-dimensional data. The methods for bar charts presented by Hlawatsch et al. [93], Borgo et al. [23] and Höhn et al. [94] are not applicable for our use case. They provide inspiration for our work – the separate representation of the exponent and the mantissa.

**Taxonomy of Tasks** Most task types can be divided into high- and low-level tasks [31]. There are several papers that deal with low-level tasks and introduce their own taxonomies of tasks [10, 146, 161, 187]. In our user study, we use two low-level tasks that appear in all of the taxonomies to evaluate the color scales.

## 3. Color Scheme Design

We present a novel color scheme named *order of magnitude colors* (Figure 2). This design is inspired by the scientific notation of numbers and is created in the *CCC*-tool [131]. Inspired by the idea of the recent approaches [23, 93, 94], we use the two parts mantissa $m$ and exponent $e$ of a value $v$ (so that $v = m \cdot 10^e$) for color coding. Every exponent given in the data is mapped to another hue. Within an exponent, the mantissa is mapped to a perceptually linear sequential scale of the respective hue.

During the research process, we created and tried many different color scales to the meteorological data (see the supplementary material). The color selection as well as the data and the study tasks were discussed and agreed upon in several focus groups organized together with the meteorologists. In the final design, the colors were chosen through a manual process so that the color distances between adjacent colors are approximately even. Color blindness would be an additional criteria for further research. We made the HSV-gradient for the different hues as smooth and equal as possible (see Figure 2). To achieve a better color nameability [148, 149], we chose unique hues instead of different saturation tones of one specific hue.

Boundaries in colors are easily perceived as changes in the exponents in the data [153]. This is deemed appropriate, as our colormap describes the changes in magnitudes. It

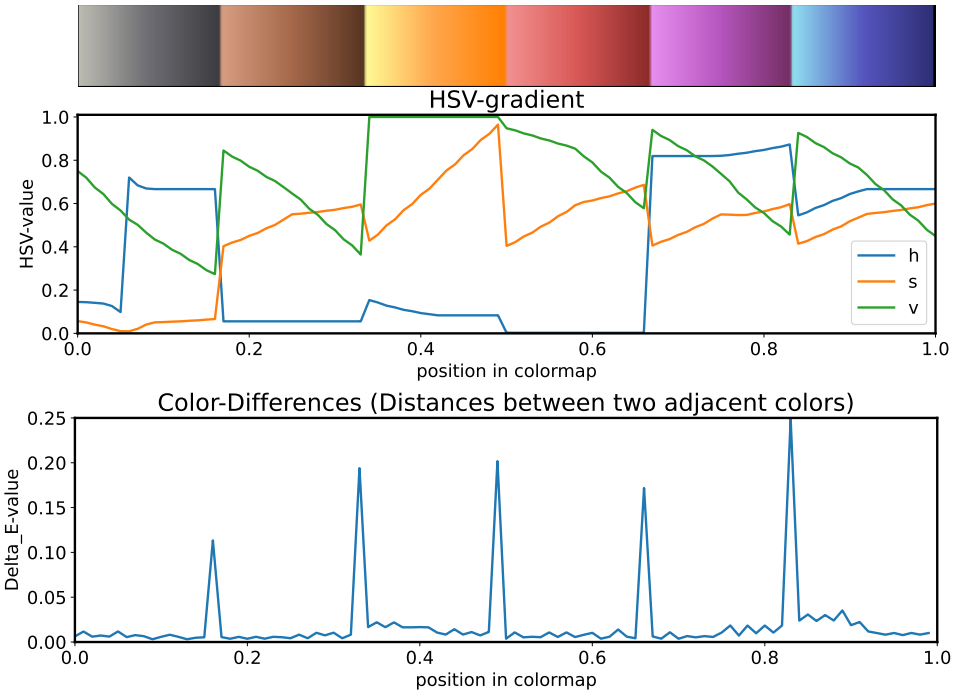



FIGURE 2: *Order of magnitude colors* color scheme.

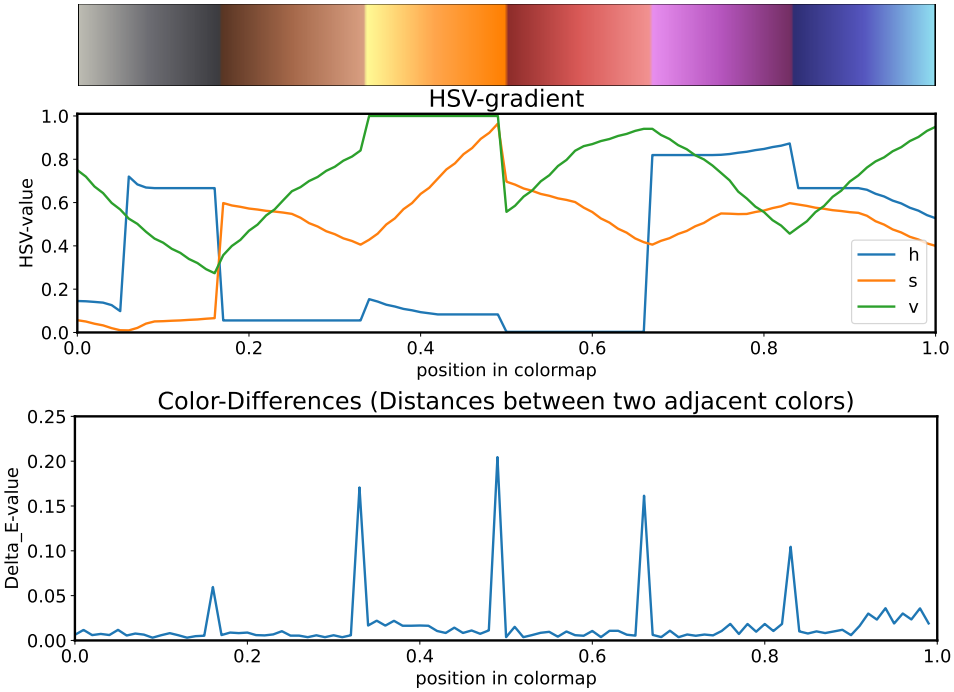



FIGURE 3: *OMC smoothed lightness* color scheme.

helps the meteorologists looking at changes in the data. Nevertheless, we also created a variation of the OMC-colormap that reduces the color distances (i.e. smaller DeltaE-values [166]) (see Figure 2 and Figure 3). In this version, we flipped every second sequential color scale so that we have a changing lightness direction comparing two neighboring color scales: The *OMC smoothed lightness* (Figure 3).

### 4. Evaluation

*Viridis* and *Rainbow* are common color schemes for visualizing meteorological data. Therefore, we evaluate these two, a state-of-the-art colormap created by *ColorCrafter* [171] and our two new designs - the OMC and $OMC_{sl}$ - in an exploratory user study.

**Data** The data used for the study were real meteorological data sets provided by the Institute for Geophysics and Meteorology at the University of Cologne. Every set of data contains one measurement day with the variables *time* (0 to 24 hours), *height* (0 to 12 kilometers) and the *ice water content* in clouds. The ice water content spans a range of $10^{-8}$ to $10^{-2}$. Due to the large value ranges, the colormaps were scaled logarithmic.

To avoid a learning effect, each question of the study contained visualizations of different measurement data. For consistency, we used data sets with similar properties, e.g., an equal value range.

**Tasks** In line with literature and after consultation with the meteorologists, we decided to focus on four main tasks, divided into reading [10,31,146,161,187] and comparison [31, 146, 161, 187] tasks:

- *Reading Tasks*:
  - *Extrema*: name the maximum and minimum exponent of a given day
  - *Value:* specify a value range for the marked region of data points
- *Comparison Tasks*:
  - *Extrema*: compare two measurement days and decide which one contains the global extremum
  - *Value:* compare the values of two marked regions of data points in one day

We manually selected the regions to ensure that the tasks for each colormap have the same level of difficulty. The comparison tasks are single choice, multiple options questions. In the reading tasks, the participants had to insert their answered numbers manually. Additionally, we asked the participants about their confidence in the given answers of the reading tasks.

**Experimental Setting and Procedure** A total of 53 participants (36 male, 13 female, and 4 prefer not to say) took part in the study. The age was distributed from 20 up to 60, but the majority of the participants were between 20 and 40 years (88%). We filtered out 5 participants, who did not correctly answer to the Ishihara tests for color blindness [48] (this was necessary because the rainbow scale is not suitable for color blindness).

The expertise of the participants was broad and ranged from a degree in mathematics (25%) to a degree in physics (13%) to a degree in computer science (10%) and others (4%), with the majority of the participants having a degree in meteorology (48%).

The study was conducted online and was set up with *LimeSurvey* [115]. There was one task per page and the participants had to click on a button manually to get to the next page, so we could store the given answers and the response time per task. The visualizations used had a resolution of 1291$x$500px. All of the participants used a computer screen with a size of at least 13".

The processing time of the study was approximately 30 minutes. All participants had to solve the same tasks. There was one task for every of the five designs for every type of task, i.e. in total we had $5[colorscales] \times 4[tasks] = 20$ trials.

After a short training phase introducing the visualization type and the types of tasks, the sorting of the tasks was randomized to reduce a learning effect. At the end of the study the participants were asked to give feedback. Study documentation is in annex.

### 4.1 Analysis

The analysis of the results is performed in *R* [147]. For every task of the study, we measure accuracy and response time. In the reading tasks, the specified confidence of the participants is also examined (on a Likert scale from 1 – very unsure to 5 – very confident).

**Correctness measurement** For every task, we compare the number of correct answers to the number of incorrect answers. In the extrema reading task, both exponents have to be correct to be considered as a correct answer. For value reading an answer is correct, if the values of the marked region lie in the answered range.

In addition, the sizes of the specified ranges in the reading value task are compared. To exclude logarithmic scaling, the size of the ranges is calculated using Equation 1:

$$\text{range size} = \frac{(Exp_{max} - Exp_{min}) \cdot 10 + (Mant_{max} - Mant_{min})}{10} \tag{1}$$

**Significance Tests** To perform our analysis, a three-stage significant test for each task was used. Since we could not assume that the data is normally distributed, we first ran a Shapiro-Wilk test on the given answers and response times, with the result that none of the data is normally distributed.

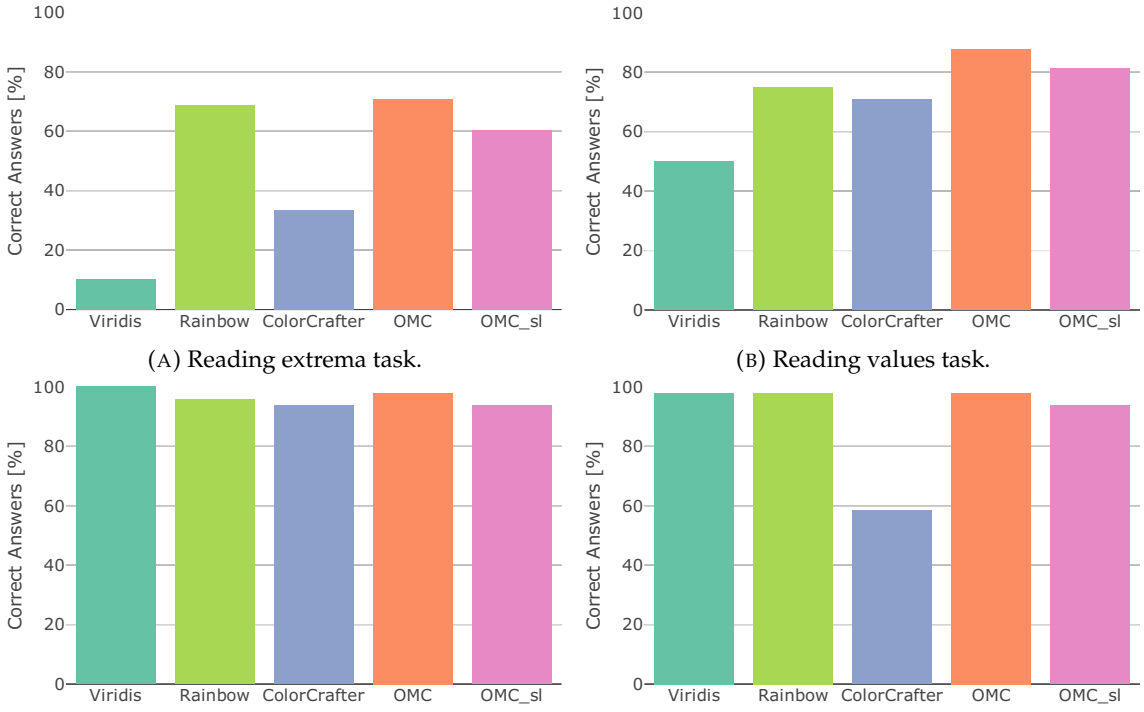
In the second stage of the analysis, we used a $\chi^2$-test (and a Fisher-test if the frequencies are too low) to investigate significance in the amounts of correct and wrong answers. For the quantitative data like the range sizes or response time, we used the Kruskal-Wallis test.

In the third stage, as post-hoc analysis we used a Wilcoxon-test respectively a $\chi^2$-test for pairwise comparison of the color schemes for tasks for which significance was found. All tests were performed with the significance level $\alpha = 0.05$.

### 4.2 Results

Figure 4 and Figure 5 show summaries of the results including the percentages of correct answers for the different colormaps in the four tasks and the response time of the participants.

**Reading Extrema** Our OMC design (71% correct answers) and the rainbow color scheme (69%) perform best. The $\chi^2$-test shows a significant dependence between the color scales and the amount of correct/ false answers ($\chi^2 = 52.466$, p-value $< 0.001$). The post-hoc pairwise test confirms that OMC and Rainbow deliver significantly better results than ColorCrafter (33%) and Viridis (10%) (Table 1). $OMC_{sl}$ is third best with 60% of correct answers.



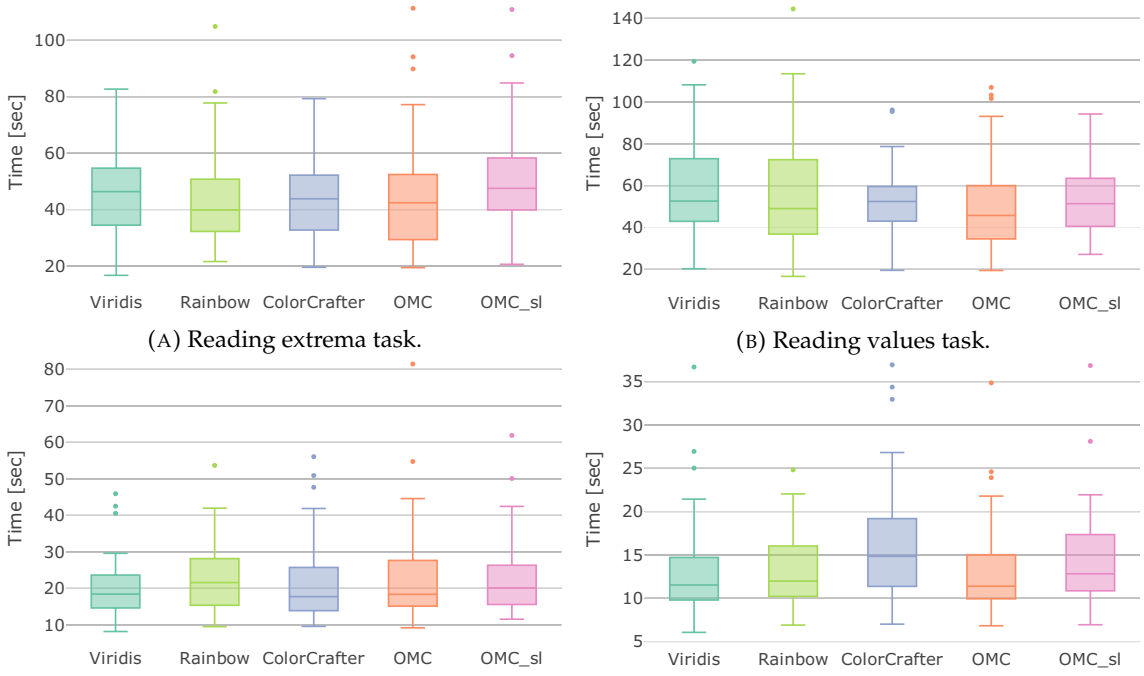

(A) Reading extrema task.

(B) Reading values task.

(C) Comparison of extrema task.

(D) Comparison of values task.

FIGURE 4: Percentages of correct answers (the more the better) for tasks and colormaps.




(A) Reading extrema task.

(B) Reading values task.

(C) Comparison of extrema task.

(D) Comparison of values task.

FIGURE 5: Response times in seconds (the less the better) for tasks and colormaps.

| **p-value** | *OMC* | $OMC_{sl}$ | *ColorCrafter* | *Rainbow* |
|---|---|---|---|---|
| $OMC_{sl}$ | 0.390 | - | - | - |
| *ColorCrafter* | 0.001 | 0.014 | - | - |
| *Rainbow* | 1.000 | 0.522 | 0.001 | - |
| *Viridis* | <0.001 | <0.001 | 0.014 | <0.001 |

TABLE 1: p-values of the pairwise $\chi^2$-test for extrema reading.

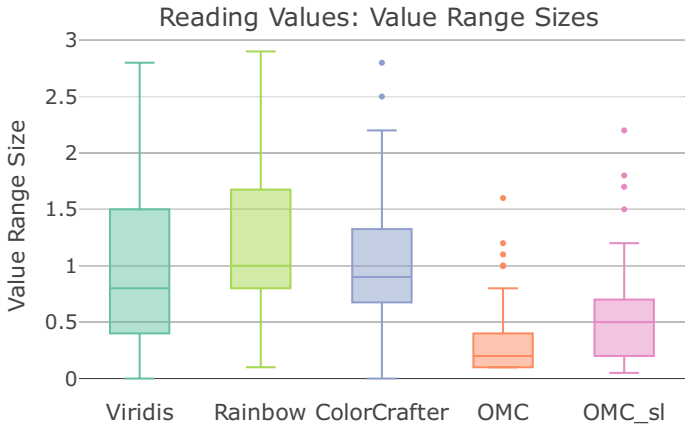



FIGURE 6: Sizes of the value ranges answered for value reading.

**Reading Values** The OMC designs perform best with respect to the number of correct answers. The $\chi^2$-test indicates significance ($\chi^2$ = 19.833, p-value = 0.001). Due to the pairwise test, only the Viridis colormap is significantly worse than other color schemes.

The positive effect of the OMC colormap can be seen better in the comparison of the given range sizes (Figure 6). The OMC colormap has by far the smallest range size (mean = 0.36), followed by $OMC_{sl}$ (0.68). Viridis (0.91), ColorCrafter (0.99) and Rainbow (1.15) are at a similar level. The Kruskal-Wallis-test confirms the significant differences in the means ($\chi^2$ = 50.518, p-value < 0.001). The pairwise Wilcoxon-test shows, that the OMC color scale leads to significantly smaller range sizes (i.e. a better identification of the searched values) than all others (Table 2). $OMC_{sl}$ also performs significantly better than ColorCrafter and Rainbow. This can be attributed to our design approach, where different hues enable fast recognizing of magnitudes.

**Comparison of Extrema** Very few errors were made. ColorCrafter (3 errors) and $OMC_{sl}$ (3) have the most errors, followed by Rainbow (2) and OMC (1). Viridis even produced no error at all. Due to the low number of errors, no significant dependence can be determined using Fisher's exact test (p-value = 0.471).

| **p-value** | *OMC* | $OMC_{sl}$ | *ColorCrafter* | *Rainbow* |
|---|---|---|---|---|
| $OMC_{sl}$ | <0.001 | - | - | - |
| *ColorCrafter* | <0.001 | 0.003 | - | - |
| *Rainbow* | <0.001 | <0.001 | 0.166 | - |
| *Viridis* | <0.001 | 0.051 | 0.511 | 0.063 |

TABLE 2: p-values of the pairwise Wilcoxon-test for value reading.

| **p-value** | *OMC* | $OMC_{sl}$ | *ColorCrafter* | *Rainbow* |
|---|---|---|---|---|
| $OMC_{sl}$ | 0.617 | - | - | - |
| *ColorCrafter* | <0.001 | <0.001 | - | - |
| *Rainbow* | 1.000 | 0.617 | <0.001 | - |
| *Viridis* | 1.000 | 0.617 | <0.001 | 1.000 |

TABLE 3: p-values of the pairwise Fisher-test for value comparison.

**Comparison of Values** ColorCrafter performs worst for this task (58% correct answers). $OMC_{sl}$ has 94% of correct answers, while the other three even have 98%. Due to the $\chi^2$-test there is a significant effect ($\chi^2$ = 66.679, p-value = 1.332e−9). The post-hoc pairwise Fisher-test shows that ColorCrafter leads to significantly more errors than all other color scales (Table 3).

### 4.3 Extended Analysis and Results

**Confidence** For the reading extrema task, the analysis of the participants' confidence shows that the sequential color schemes (Viridis, ColorCrafter) lead to significantly more uncertainty in the answers than the multicolored schemes (OMC, $OMC_{sl}$, Rainbow).

For reading value task, the participants were significantly more confident with our OMC color scale than with all others, both for the exponents and the mantissas of the specified range.

**Time** Only for the value comparison task, the Kruskal-Wallis-test shows a significant mean effect in the response time ($\chi^2$ = 13.398, p-value = 0.009). As the pairwise Wilcoxon-test shows, participants used significantly more time with ColorCrafter than with all others. Notably, the participants did not need any additional time to understand our new *order of magnitude colors* design.

**Free Text** 31 participants gave free text feedback. Almost all of them were positive regarding our OMC color scale. We already received requests to use our design. It was stated that the color scheme supports value comparison and identification by its clear borders: "The color scale with different colors between every power of 10 is the one that allowed me to better identify the values and the differences between points." Suggestions for improvement were the color selection and the perception of qualitative patterns.

The sequential colormaps were perceived as more aesthetically pleasing. Rainbow and ColorCrafter got the most negative feedback, especially for their color gradients.

**Expertise** Comparing the results of participants with a meteorological background to the others, there was no significant difference in the correctness of their answers and in the response time. This indicates that our color scheme can be used broadly, without specific expertise or background.

**Limitations** There are some limitations to our color scheme. The approach is not suitable for comparing values that are close to the borders of the exponents or for data containing positive and negative values. Additionally, further experiment may be needed to assess the scalability of the proposed scheme for value ranges with different exponent distributions. In order not to make the user study too extensive, we focused on a subset of tasks most relevant to our collaborators. Therefore, some types of task could not be

dealt with sufficiency. For example, an evaluation of the OMC colormap for high level tasks would be interesting.

## 5. Conclusion and Future Work

We presented a new color coding approach to visualize data featuring large value ranges with the application to meteorological data. The empirical study has shown that our order of magnitude colors design performs very well for all tasks examined. It has no significant difference to the best performing colormaps in the comparison tasks and improves the accuracy of value identification significantly. The strict borders of our color scheme support the perception of the different orders of magnitude.

In summary, our results suggest the following ranking for each proposed design:

- *Reading Extrema*:
  **OMC** $\succeq$ Rainbow $\succeq$ OMC$_{sl}$ $\succ$ ColorCrafter $\succ$ Viridis
- *Reading Values*:
  **OMC** $\succ$ OMC$_{sl}$ $\succeq$ Viridis $\succeq$ ColorCrafter $\succeq$ Rainbow
- *Comparison of Extrema*:
  Viridis $\succeq$ **OMC** $\succeq$ Rainbow $\succeq$ OMC$_{sl}$ $\sim$ ColorCrafter
- *Comparison of Values*:
  **OMC** $\sim$ Viridis $\sim$ Rainbow $\succeq$ OMC$_{sl}$ $\succ$ ColorCrafter

In general, our study results confirm findings from Golbiowska et al. [83] that multi-colored scales like OMC and Rainbow are more suitable for reading tasks while sequential scales like Viridis work better for comparing tasks.

Our design was developed on the application example of multidimensional meteorological cloud data. Our evaluation was data agnostic making our results extendable to other data sets. The choice of color scheme however was informed by our collaboration with domain expert therefore, in the future, we would like to test the performance of our approach on other data types from various application areas, potentially exploring new color schemes. Furthermore, we have to investigate color blindness for our new design.

## Acknowledgments

The authors would like to thank all study participants and the reviewers, whose suggestions helped improve this paper.

## 2.2 Novel Visualization Designs for Time-Series Data with Large Value Ranges

In this work, two novel visualization designs are developed that extend existing techniques to highlight the specific characteristics and requirements of large value ranges in individual time-series. One of the designs uses the OMC color scale developed in the previous paper to support the perception of magnitude variations.

The paper is published in IEEE TVCG and was presented at the IEEE Visualization conference:

**D. Braun**, R. Borgo, M. Sondag, and T. von Landesberger. Reclaiming the horizon: Novel visualization designs for time-series data with large value ranges. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1161–1171, 2024.
doi: 10.1109/TVCG.2023.3326576

The supplementary material of the paper, including the study results and documentation as well as the Python code for the data and design generation, is publicly available at OSF.

I am the primary author of this publication. In this role, I was responsible for the design, implementation, data collection and analysis, as well as the writing and publication of the work. The specific contributions of myself and my co-authors to this publication are outlined below according to the Contributor Roles Taxonomy (CRediT):

**D. Braun**: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Software, Visualization, Writing – Original Draft. *M. Sondag*: Visualization, Writing – review & editing. *R. Borgo, T. von Landesberger*: Supervision, Conceptualization, Methodology, Writing – review & editing.

# Reclaiming the Horizon: Novel Visualization Designs for Time-Series Data with Large Value Ranges

DANIEL BRAUN[1], RITA BORGO[2], MAX SONDAG[1], TATIANA VON LANDESBERGER[1]

[1]*University of Cologne*
[2]*King's College London*

**Abstract:**

We introduce two novel visualization designs to support practitioners in performing identification and discrimination tasks on large value ranges (i.e., several orders of magnitude) in time-series data: (1) The *order of magnitude horizon* graph, which extends the classic horizon graph; and (2) the *order of magnitude line* chart, which adapts the log-line chart. These new visualization designs visualize large value ranges by explicitly splitting the mantissa $m$ and exponent $e$ of a value $v = m \cdot 10^e$. We evaluate our novel designs against the most relevant state-of-the-art visualizations in an empirical user study. It focuses on four main tasks commonly employed in the analysis of time-series and large value ranges visualization: identification, discrimination, estimation, and trend detection. For each task we analyze error, confidence, and response time. The new *order of magnitude horizon* graph performs better or equal to all other designs in identification, discrimination, and estimation tasks. Only for trend detection tasks, the more traditional horizon graphs reported better performance. Our results are domain-independent, only requiring time-series data with large value ranges.

*IEEE Transactions on Visualization and Computer Graphics, 2024*

## 1. Introduction

Large value ranges (i.e., several orders of magnitude) in time-series data are common to a wide range of application domains. These large variations in orders of magnitude can occur in a relatively short amount of time in a variety of application domains. Examples include medicine with pandemic outbreaks (e.g., COVID-19 cases in Germany, which range from 10 000 to 300 000 per day within 100 days in Figure 1), meteorology with measurements of storms (e.g., the ice water content in clouds, which range from $10^{-7}$ to $10^{-3} kg/m^3$ in hours [29]), or finance with stock markets (e.g., bitcoin cryptocurrency with price changes of more than 40 000 dollars within a year [133]).

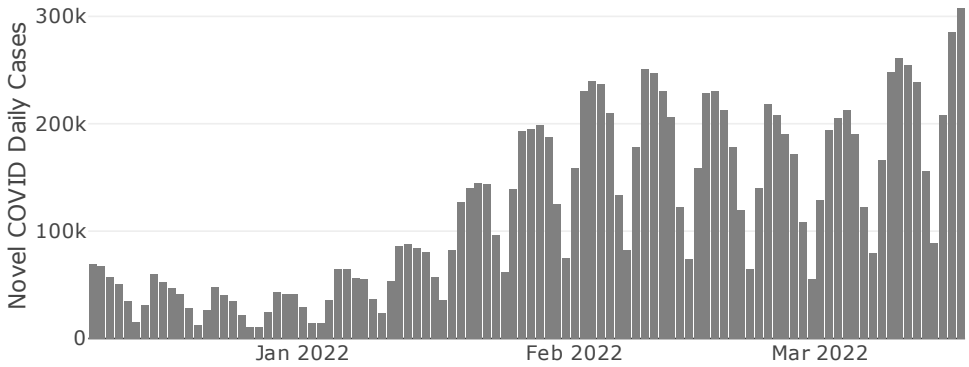



FIGURE 1: Number of daily new COVID-19 cases in Germany.

Research on visualization of short time-series has proposed techniques to improve accuracy and completion time of low-level tasks compared to standard line charts [4,68].

However, these visualizations fail in visualizing large orders of magnitude. Figure 2b shows one such technique: a horizon graph [71, 150], which divides and layers the line chart. Here, small changes in values are difficult to detect across all orders of magnitudes, and especially difficult is reading and comparing values at lower magnitudes. In Figure 1 we see another example of this. At the same time, research on visualizing large value ranges focuses mainly on uni-variate data without time [23, 93, 94]. This paper combines these two areas, visualizing large value ranges in time-series data.

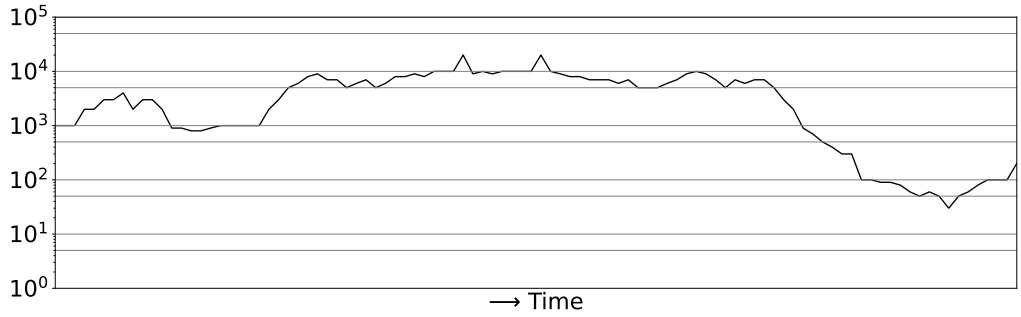
We present two novel visualization designs that improve two standard visualization techniques for time-series data – log-line charts and horizon graphs – to meet the challenges of magnitude variations. First, the *order of magnitude line (OML)* chart (Figure 2d) adapts the log-line chart by using linear mapping within each order of magnitude. In addition, the OMC color scale [29] – a color scale designed for large value ranges – is used to support the perception of magnitude variation and value identification. The second proposed design – the *order of magnitude horizon (OMH)* graph (Figure 2e) – adjusts the standard horizon graph by using a separate representation of the mantissa and exponent of each value. Every order of magnitude is represented by a separate band, while the mantissa is mapped to a linear scale on the y-axis for easier identification of individual values. In this paper, we assume that numbers are encoded in base 10, but other bases could be supported.

To evaluate our new designs, we conduct a user study that compares them with the standard visualization techniques for time-series – log-line charts (Figure 2a) and horizon graphs – as well as one state-of-the-art approach for large value ranges for time-independent data – the scale-stack bar chart [93]. Our study focuses on low-level visualization tasks from taxonomies of both time-series [4, 5] and large value ranges [23, 29, 93, 94]. The results show that OMH outperforms their currently used counterparts in tasks of "identification" (i.e., read a marked value), "discrimination" (i.e., compare two marked values in one visualization), and "estimation" (i.e., determine the difference of two marked values). OMH has both the lowest error rate and the highest confidence of all designs in these three tasks, while the time taken by the user is similar to the others. For OML, we only find a significant effect on confidence when compared to horizon graphs. The standard horizon graph on "detect trend" (i.e., identify the trend of the visualized data), tasks proves to be better than all designs tested except OMH, likely due to the linear scaled y-axis as opposed to logarithmic scaling.

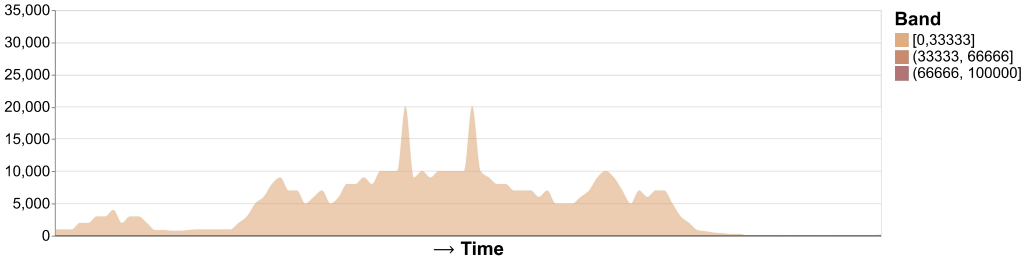
The primary **contributions** of our work are:

- We introduce *two novel visualization designs*, extending existing time-series designs to meet the requirements of data with large value ranges.
- We *empirically evaluate* our new designs with state-of-the-art designs in a user study with 90 participants.
- The *novel designs are domain-independent* and can hence be used on a variety of application domains (e.g., meteorology, finance, healthcare) to better present temporal data with large value ranges.

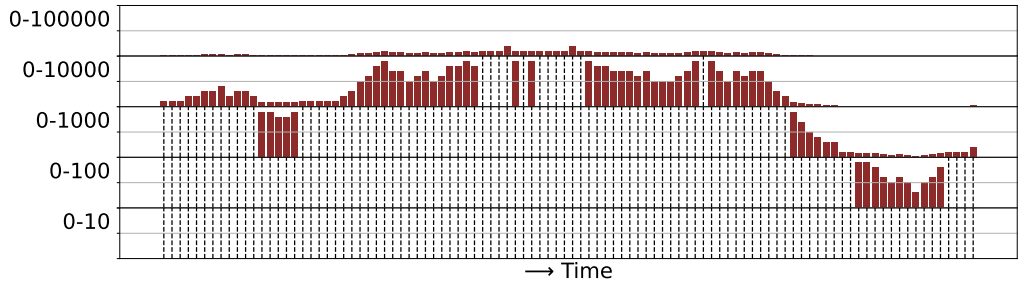
After an overview of current methods and evaluations in Section 2, we introduce our two new designs in Section 3. The design of our user study is presented in Section 4 and its analysis and results are described in Section 5. In addition to a discussion of the findings in Section 6, we also point out limitations and future work in Section 7.



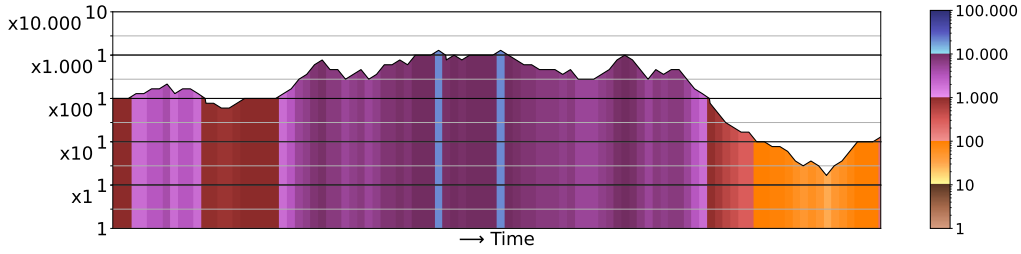

(A) Log-line chart



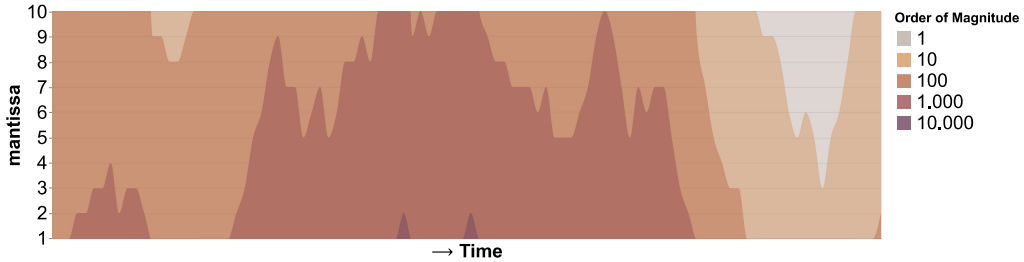

(B) Classic horizon graph [71, 150]




(C) Scale-stack bar chart (SSB) [93]




(D) Order of magnitude line chart (OML)




(E) Order of magnitude horizon graph (OMH)

FIGURE 2: Exemplar time-series data with large value ranges encoded using three standard visualization types log-line chart (a), horizon graph (b), scale-stack bar chart (c), and our two new designs *order of magnitude line* (d) and *order of magnitude horizon* (e).

## 2. Related Work

We present the latest solutions and evaluations for displaying time-series as well as for visualizing data with varying orders of magnitude.

**Visualization of Time-Series** The visualization of time-series data has been widely investigated using a variety of different visual designs [2, 4, 5, 68]. The most popular way to present continuous coherence in time-series is the line chart [8, 51]. Over time, several novel visualization techniques have been designed and evaluated in relation to the traditional line graph. The combination of quantitative data with qualitative abstraction (e.g., a composite representation with color encoding of qualitative data and spatial position encoding of the quantitative data) has been studied by Aigner et al. [7] and Federico et al. [69]. We adapt the idea of adding qualitative coloring based on a division of the value range to a line graph in our *order of magnitude line* design. Meaningful color coding is used successfully in other time-series visualizations, such as ThemeRivers [87], to show thematic changes over time, and RankExplorer to visualize ranking changes over time [168].

In recent years, the horizon graph has become increasingly popular. The final design was presented for the first time by Reijner [150] and Few [71], while the principle of a two-tone pseudo coloring was already used some years before by Saito et al. [160]. In a horizon graph, the value range is divided into several bands of the same size, which are then superimposed onto each other to create a layered form. Color (hue and saturation) is used to distinguish the individual bands. Heer et al. [90] compared the horizon graph with line graphs and investigated the appropriate number of bands. They found that horizon graphs improve estimation accuracy and using more than three bands increases the error and time. Therefore, we use three bands for the standard horizon graph in our evaluation. Jabbari et al. [98, 99] compare the horizon graph to alternative composite visual mappings, such as a hue-saturation mapping which is very similar to warming stripes [88]. They found that the horizon graph performed best in terms of discrimination and estimation errors, and was only slightly slower than the proposed alternative mappings. For this reason, we have decided not to evaluate a warming stripes variant for large value ranges in our study. Gogolou et al. [81] explore the aspect in comparing similarity perception between line charts, horizon graphs, and colorfields. The horizon graph promotes local variations the most, while the other two have advantages for amplitude and y-offset scaling.

One of the advantages of the horizon graph is the space saved by layering the different bands, which makes it especially efficient for multiple time-series. Although we do not focus on multiple time-series in this paper, the results of Javed et al. [100] for horizon graphs in that specific case are interesting. They show faster perception in discrimination tasks for horizon graphs than for simple line graphs and a similar correctness rate compared to the other evaluated designs for every task. The study of Perin et al. [141] supports these results. Here, multiple horizon graphs outperform multiple line charts in each task, with even better results for an interactive version of the horizon graphs. These studies indicate that extending our OMH to multiple time-series could prove promising in future work.

**Visualization of Large Value Ranges** A common way to display time-series with order of magnitude variations is to use a line graph with logarithmic scaling [111, 155, 164]. However, most of the previous research on large value ranges has been done for time-independent data.

Hlawatsch et al. [93] were the first to develop a new visualization technique for displaying large-value ranges and adapted the classic bar chart. Their *scale-stack bar chart (SSB)* (Figure 2c) represents each value at multiple stacked scales with increasing value ranges. Each scale starts at zero and maps the numbers linearly. The results of a user study showed improvements of their approach compared to linear and logarithmic bar charts for quantitative comparisons.

Borgo et al. [23] use the separate representation of exponent and mantissa by showing both parts of each value as two overlapping bars on one linear scale in a classic bar chart, and call them *order of magnitude markers*. Their evaluation confirms that designs that consider large value ranges increase the accuracy for evaluation tasks.

By encoding the orders of magnitude using the color and width of classic bar charts, Höhn et al. [94] adapted the previous method. They compare their *width-scale bar charts* to the two other approaches in a quantitative study. The results showed that their design works best for interpretation tasks. However, for estimation, discrimination, and trend tasks, the SSB resulted in the highest accuracy of the three approaches. Therefore, we decided to include the SSB in our evaluation.

Braun et al. [29] showed that color can be used to support the perception of changes in the orders of magnitude. Their *order of magnitude color (OMC)* colormap encodes a value $v = m \cdot 10^e$ by mapping the exponents $e$ to the hue and the mantissa $m$ to a sequential color scale of that hue. We use this color scheme in our OML design.

## 3. Visualization Designs

The presented visualization designs combine methods for time-series data as well as methods for data with large value ranges. We designed our visualizations as static, non-interactive visualizations. While interactivity could help in understanding charts with multiple series [68, 141] or long time-series [4, 5], we focus on short and singular time-series here. Moreover, the static views can be used in printed media.

The "default" way to visualize large value ranges, is to use a logarithmic scale. However, reading values from a logarithmic scale is not intuitive [111]. This is evidenced by the focus of previous techniques for large value ranges on perceptually linear scaling for both the mantissa $m$ and the exponent $e$ of a value $v = m \cdot 10^e$ [23, 93, 94]. Our approach reflects techniques for large value ranges by splitting the mantissa $m$ from the exponent $e$ and treating them differently. Values within the same exponent $e$ are visualized continuously, while there is a jump between values with a different exponent. This continuity is an advantage over the SSB approach, where a coherent reading of the data is interrupted by the need to repeatedly restart the scales at $v = 0$.

### 3.1 Order of Magnitude Horizon

The *order of magnitude horizon (OMH)* graph is an adaptation of the standard horizon graph to meet the requirements of data with large value ranges. Figure 3 illustrates the construction of OMH. Unlike the standard horizon graph, the starting point is a log-line graph (Figure 3a). After scaling the y-axis linearly within each order of magnitude (Figure 3b), the graph is split into uniformly-sized bands. Each band represents a different order of magnitude, and hence the number of bands depends on the number of orders of magnitude. We assume that numbers are encoded in base 10, but other bases could be supported. The colors of the bands are based on the hue gradient of the OMC color

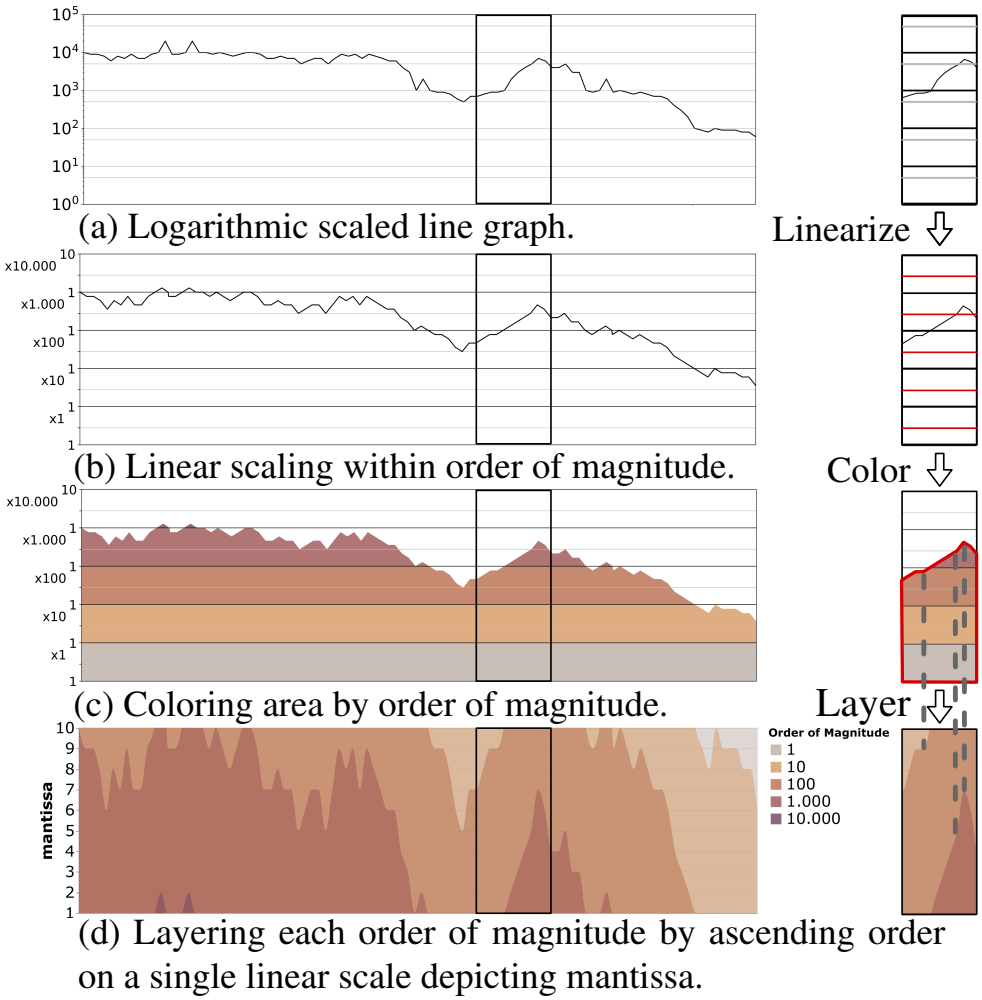



FIGURE 3: *Order of magnitude horizon* graph: Step-by-step construction.

scale [29]: The larger the magnitude of the band, the more saturated its color (Figure 3c). Finally, the different bands are layered on top of each other, sorted in ascending order (Figure 3d). By using two different visual variables, vertical position for the mantissa $m$, and saturation for the exponent $e$, both $m$ and $e$ can be directly compared in our design. Even though not evaluated, it should be possible to use OMH for data with both positive and negative values with a diverging color scale.

### 3.2 Order of Magnitude Line

The order of magnitude line (OML) graph is an adaption of the log-line graph. Starting off with a log-line graph (Figure 4a), the y-axis is scaled linear within each order of magnitude (Figure 4b) to facilitate easier reading of the values. In addition, the area below the line in the graph is colored with the OMC [29] color scale (Figure 4c) to further improve the perception of the values. Hence, the values are double encoded by the visual variables vertical position and color. In the OMC color scale, each exponent $e$ is given a different hue, with a sequential color scale for the mantissa $m$. This change in hue between the different exponents $e$ reduces the probability of a magnitude error.

### 3.3 Design Novelty

Compared to traditional methods, both OML and OMH designs propose a novel approach with respect to spatial organization of the data. Both designs enforce continuity

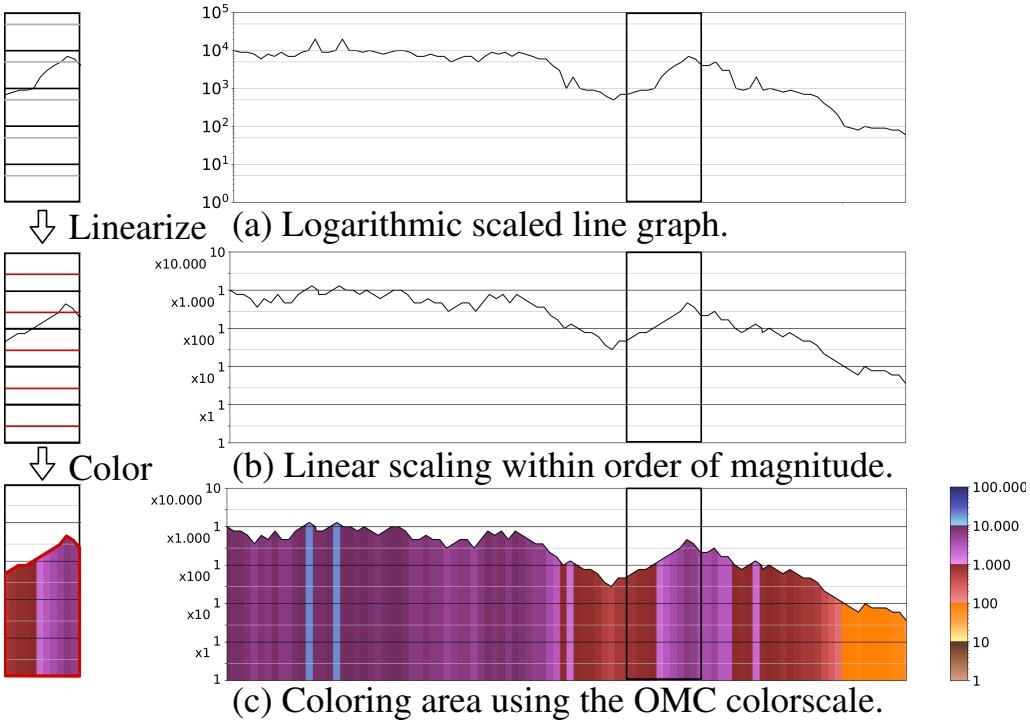



FIGURE 4: *Order of magnitude line* graph: Step-by-step construction.

in the visual layout. This is lost in traditional approaches dealing with visualization of large magnitude values, where priority in the design is given to the enhancement of value variations. Our designs instead intend to favour perceptual principles, leveraging Gestalt [194] principles of grouping through proximity, similarity, continuity and common faith.

### 4. Evaluation

To evaluate our two novel approaches, we compare them to their standard counterparts (Figure 2) – log-line chart and horizon graph for time-series data as well as SSB charts for data with large value ranges – in an empirical user study to test the following hypotheses:

- **H1:** *Our new designs reduce error rates in value identification, discrimination, estimation, and trend detection tasks on data with large value ranges compared to their standard counterparts.*
  As our designs are explicitly designed based on research for large value ranges, which have been proven to improve performance on these tasks [23, 29, 93, 94], we expect this improvement to translate towards time-series with large value ranges.

- **H2:** *Our new designs increase participants' confidence in their answers compared to their standard counterparts.*
  We expect that the linear scaling of OMH and OML makes them easier to read. Furthermore, participants have multiple options to confirm their answers due to the double encoding by color.

- **H3:** *Our new designs have higher response times than the standard log-line chart and comparable response times to the others.*
  Participants should be familiar with standard log-line charts. Due to the novelty, we expect that they will need more time to interpret our designs.

### 4.1 Stimuli Design

Beyond the question of which visualization techniques to evaluate, there are additional ways to improve charts to allow better interpretation of presented data. Throughout the stimuli design process we followed state-of-the-art visualization guidelines [61, 129, 193].

Considering which visual elements to include or omit in the stimuli designs quickly leads to the question of "chartjunk" [72]. Research on this topic shows that visual embellishments that are not essential to understanding the data can help to transmit the story of a visualization, but distract from the quantitative data [16]. The goal of our user study is to test low-level tasks such as reading and comparing values correctly, rather than whether aesthetically pleasing designs perform better or worse. That's why we use minimalistic presentations and treat each design in a similar way.

To ensure comparability in the evaluation, we fixed some design criteria and applied them consistently to all five visualizations (Figure 2). The three charts with a color coding – classic horizon graph, OMH, and OML – all include a color legend to increase legibility. Similar to previous studies on visualizing time-series [81, 98, 99, 141], we have chosen not to show the time point labels on the x-axis to avoid semantic meaning that could distract from the actual study tasks. Y-axis tickmarks are displayed on the main grid lines. Previous research on grid lines shows that, on the one hand, they support the perception of individual values, on the other hand, they can be overwhelming and distracting, which is why they should be used cautiously and not too obtrusively [15,70,175]. Therefore, we decided to include only one grid line indicating the mantissa value 5 in addition to the main grid lines used to distinguish the orders of magnitude. The lines within the sections have been drawn with less opacity so that they do not interfere with the visualization.

An additional element that can interfere with the perception of the visualizations are the visual markers used to highlight the data point of interest in the study tasks. We tried different types of visual markers such as arrows or vertical lines inside the visualizations, but found that these distracted from the visualization of the data. Therefore, we use a visual marker outside the visualization. To create a comparable environment to the previous horizon graphs studies [90, 98, 99], we used the same marker design. The data points are highlighted by two small vertical lines at the top and the bottom of the visualization, as well as a letter above the upper line (see Figure 5).

All charts shown in the study had a size of 972x350px to ensure that the visualizations are fully visible on the display for all participants without scrolling. Since we do not consider multiple time-series, each trial contains exactly one visualization.

### 4.2 Data

In comparable studies on large value ranges [23, 93, 94], the data covered a range of $[0, 10\,000]$ with only integer mantissa. We extend this value range to a maximum value of 100 000, so that our data can be described by $m \cdot 10^e$ where $1 \leq m \leq 10$, $0 \leq e \leq 4$ and $m, e \in \mathbb{Z}$. As we do not focus on streaming applications or long time-series, each data set consists of 100 time steps for consistency.

We used synthetic data sets generated by a constrained random walk. A uniform random value between 1 000 and 10 000 is the starting point of the walk, from which the mantissa is changed by a uniform random value between $[-2, 2]$ at each step. Smoothed generated walks might cover only a fraction of time-series that occur in reality, but they help us to control the study conditions and ensure an equal level of difficulty for all trials. Moreover, they are already used for data generation in previous studies on horizon graphs [90, 98, 99].

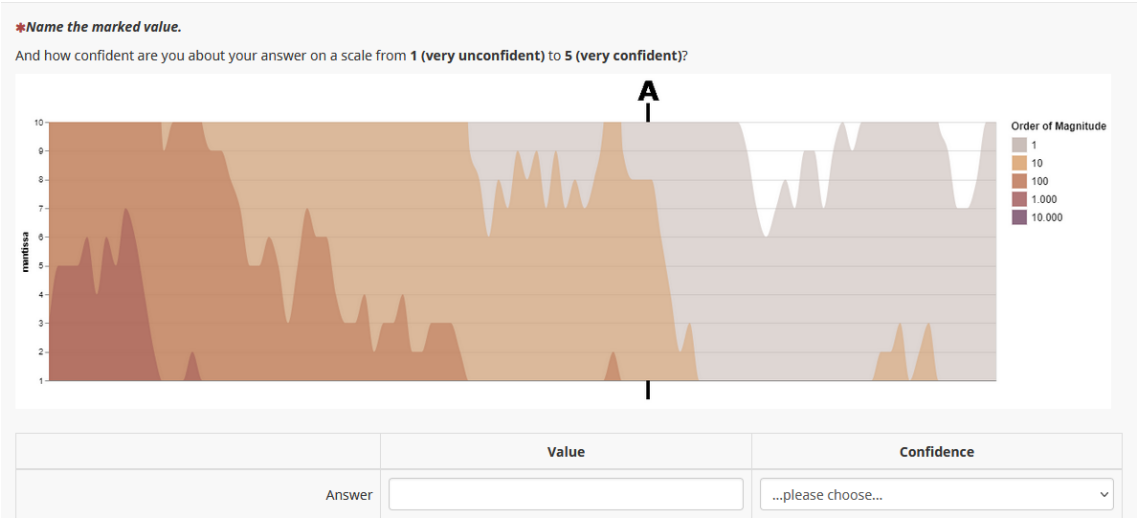



FIGURE 5: The interface of the user study. This example shows a reading task for the OMH design. The participants type a numerical answer for value *A* in a text box and indicate their confidence level on a drop-down Likert scale.

### 4.3 Tasks

In line with state-of-the-art task taxonomies, we focus on four low-level tasks [10, 31, 187] corresponding to our hypothesis H1. The selected tasks are a combination of common tasks for both areas we relate to in this paper – the visualization of time-series data and data with large value ranges. Discrimination and estimation tasks are mostly examined in studies on time-series visualizations [7, 90, 98–100, 108, 141], while identification and trend analysis tasks can be found in both fields [7, 23, 29, 69, 93, 94, 108, 190].

In order to ensure that the trials were comparable and to provide a larger set of answers for the analysis, we split each task into three conditions. The marked values are selected from the previously generated data sets depending on these conditions. We reduced potential bias from the position of these marked values by randomization and multiple trials (three trials per task with different data conditions). The conditions per task we are investigating are as follows:

**Identification – *Value Reading*** The aim of the "Identification" task is to read individual values as accurately as possible. In each trial, the participant is shown a data set with a marker at a particular value. The target data point is randomly selected with constraints from these conditions:

- Condition 1: The value is on a grid line.
- Condition 2: The value is in a high order of magnitude ($10^3$ or $10^4$) and not on a grid line.
- Condition 3: The value is in a low order of magnitude ($10^1$ or $10^2$) and not on a grid line.

The participants enter their identified value manually via a text box. The interface of an example identification question is shown in Figure 5.

**Discrimination – *Value Comparison*** The aim of the "Discrimination" task is to compare two values and determine which one is larger. As we are not looking at multiple time-series, the discrimination between the two data points takes place in one visualization. In each trial, the participant is shown a data set with two markers at different values. The data points are marked with $A$ and $B$, with $A < B$ on the x-axis. The selection of data points is random, with constraints imposed by the following conditions:

- Condition 1: The values are in the same order of magnitude.
- Condition 2: The values are in neighbouring orders of magnitude.
- Condition 3: The values are in distinct orders of magnitude with a difference of at least one order of magnitude.

The participants select the letter of the larger value from a drop-down menu.

**Estimation – *Difference Determination*** The estimation task goes one step further with regard to the discrimination task. The aim here is to determine the absolute, quantitative difference between two values in one time-series. Structurally, the task is similar to the discrimination task. Again, the data points are selected randomly according to the same conditions and are also labeled A and B:

- Condition 1: The values are in the same order of magnitude.
- Condition 2: The values are in neighbouring orders of magnitude.
- Condition 3: The values are in distinct orders of magnitude with a difference of at least one order of magnitude.

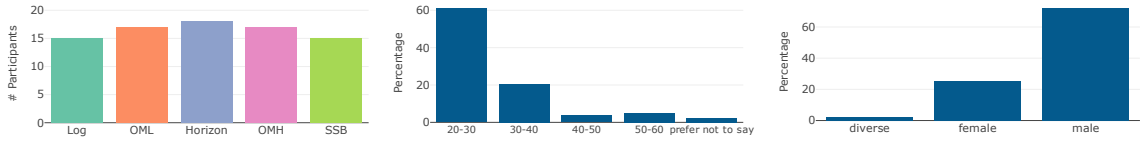
As in the identification task, the participants enter their answers manually via a text box.

**Trend – *Trend Detection*** The aim of the "Trend" task is to identify the trend in the visualized data. One visualization without any marker is shown for each trial. For this task, we do not use data generated from a random walk described in Subsection 4.2. Instead, we created pseudo-random data to simulate specific types of trends. Pseudo-random in this case means that the order of the magnitudes is predetermined to generate a specific kind of trend, but the values for the mantissas are chosen randomly. Although they certainly do not reflect all common trends, we have chosen to test the following three types, based on the selection made in the comparable study task by Höhn et al. [94]:

- Condition 1: Periodic trend.
- Condition 2: Linear trend.
- Condition 3: Exponential trend.

The participants select the presented trend from a drop-down list containing the options *periodic*, *linear*, *exponential*, and *none*.

In addition to these tasks, we ask the participants about their confidence in their answers for each trial using a 5-point Likert scale [186]. We use this as an indicator of the difficulty of interpreting the visualization as perceived by the participants.




(A) Number of participants per design.

(B) Distribution of participants per age (in %).

(C) Distribution of participants per gender (in %).

FIGURE 6: Demographic characteristics of the study participants (excluding participants with color vision deficiencies).

### 4.4 Experimental Setting

**Procedure** The study was conducted online and was set up with LimeSurvey [115]. We opted for a between-subject study design [38], mainly for two reasons: First, it excludes the learning effect that would occur in a within-subject study due to the order in which the designs are processed. Second, the number of trials per participant is reduced, which allowed us to divide the tasks into the different data conditions for more expressive results (see Subsection 4.3).

In the study interface, each page contained one trial and participants had to manually click a button to move to the next page so that we could store the answers given and measure the response time per trial. A back button has been omitted to avoid learning effects. As the study was online, we were unable to control the size and properties of participants' screens, but we recommended a 13" or larger screen. The influence of confounding variables, such as different screen sizes or color vision deficiencies, was reduced as much as possible. Potential uncontrollable external influences such as distracting environmental noise was averaged out due to the large number of participants.

The average processing time of the study was about 15 minutes. Each participant had to complete three trials per task (the three data conditions), resulting in a total of $1[design] \times 4[tasks] \times 3[conditions] = 12$ trials per participant. After each task, the participants were invited to have a short break if needed. A unique data set was generated for each trial (see Subsection 4.2), so that a total of $12[trials] \times 5[designs] = 60$ different data sets were used in the study.

After demographic questions, which included single-choice questions about the participants' gender, age, degree, and experience with time-series visualizations, and a short training phase to introduce the types of tasks and input options, the participants were randomly assigned to one of the five designs. Before the actual tasks, the visualization designs were explained and various attention questions were included between the tasks. At the end of the study, the participants were asked to give free text feedback. The study documentation is presented in the supplementary material.

**Participants** A total of 90 participants (64 male, 23 female, 2 diverse, and 1 prefer not to say) took part in the study (Figure 6c). The age was distributed between 20 and 60, but the majority of the participants (81.7%) were between 20 and 40 years old (Figure 6b). As the OMC color scale is not suitable for color vision deficiencies, we use the Ishihara test for color blindness [48] before beginning the study. 8 participants did not answer the questions correctly, and were hence filtered out from the results. All participants came from a university environment, i.e., they were students or had a higher academic degree, and were recruited through advertising in lectures, mailing-lists, and word of mouth. Thus, they should be familiar with exponential notation.

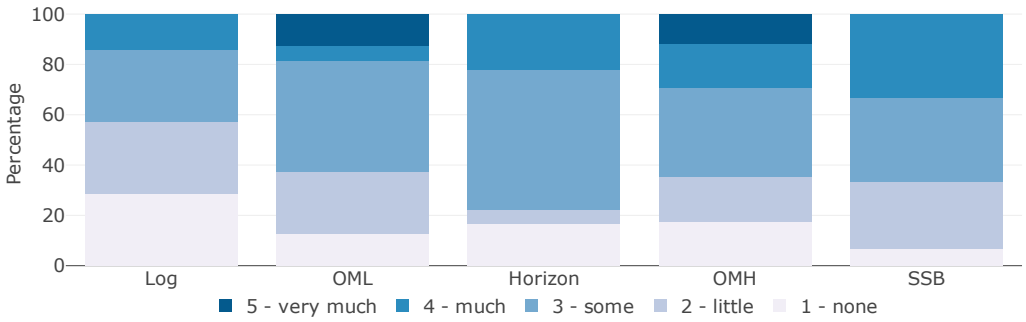



FIGURE 7: Distribution of participants' self-reported expertise per design.

The random allocation of participants among the different designs resulted in the following distribution: 15 people completed the tasks for the log-line chart, 17 for OML, 18 for the classic horizon graph, 17 for OMH, and 15 for SSB (Figure 6a). Due to the random assignment in the between-subject study, there was a potential risk of an expertise bias. We tested for dependence of the participants' reported expertise in time-series using a $\chi^2$-test, which showed no statistical evidence that designs and expertise correlate with each other ($\chi^2$ = 16.168, p-value = 0.4431). Therefore, the results are comparable (see Figure 7).

## 5. Results

In the described study setting, the visualization design as well as the tasks with their different data conditions are the independent variables. For each task, we analyze error (inaccuracy), confidence, and response time (i.e., the dependent variables of the study), and provide a ranking of the proposed designs based on the results. In addition, we summarize the participants' open-ended feedback.

### 5.1 Analysis

**Error Measurement** As we have quantitative answers for identification and estimation tasks, but categorical answers for discrimination and trend tasks, we use two different error definitions to test H1.

To allow comparison with the results of Hlawatsch et al. [93], we use the same error calculation for the identification and estimation tasks: $error = |1 - \frac{response\ value}{correct\ value}|$. This measures the relative deviation from the correct value (the lower, the better). This error definition particularly takes into account the characteristics of logarithmic scaled data. Thus, $error_1 = |1 - 10/100|$ and $error_2 = |1 - 1000/10000|$ give the same error value ($error_1 = error_2 = 0.9$), although the absolute error is different (90 resp. 9 000). Therefore, an error $\geq 1$ indicates an exponent error.

We considered the method of Borgo et al. [23] to give a 20% error tolerance to the correct value to evaluate the accuracy of an answer. However, it gives an inaccurate error value due to answers being either just correct or not. They used this method because some of their tasks resulted in uncertainty in the answers. This is not the case for us, as we ask for precisely determinable values in our study. Thus, we decided not to use this method.

For the discrimination and trend tasks, we use the error calculation of Höhn et al. [94]. Due to the single-choice nature of the questions, there is only the binary result of *true* or *false* in both tasks. By replacing true and false with 1 and 0 (representing 100% resp. 0% correctness), we create an accuracy of the answers. Subsequently, these are converted

into error values by 1 – *accuracy*. This transformation preserves that lower error values represent better results.

**Significance Tests** To analyze the results, we used a three-stage significance test for each task. We first ran a Shapiro-Wilk test on the answers and response times to test if the data is normally distributed. The test results showed that there is no normal distribution for any of the tasks. Therefore, we use non-parametric tests for further steps.

The second stage of the analysis is divided into tests for quantitative and categorical data. For the categorical confidence responses, we tested the answers for independence in the designs with the $\chi^2$-test. A dependence of confidence and design indicates a significant difference. For the quantitative errors and response times, we used Kruskal-Wallis to test the designs on different means.

As post-hoc analysis for the independent observations we used a Wilcoxon-Mann-Whitney-test for the quantitative answers and a $\chi^2$-test for the confidence answers. This was done only for tasks and answers that were found to be significant in the second stage. This allows a comparison of all design combinations and the testing of our hypotheses. The p-values of the combinational comparisons are presented in Section 5 in the form of triangle matrices.

All test were performed with the standard significance level $\alpha = 0.05$ and a Bonferroni correction factor of 10, corresponding to the number of pairwise comparisons per task and measured aspect.

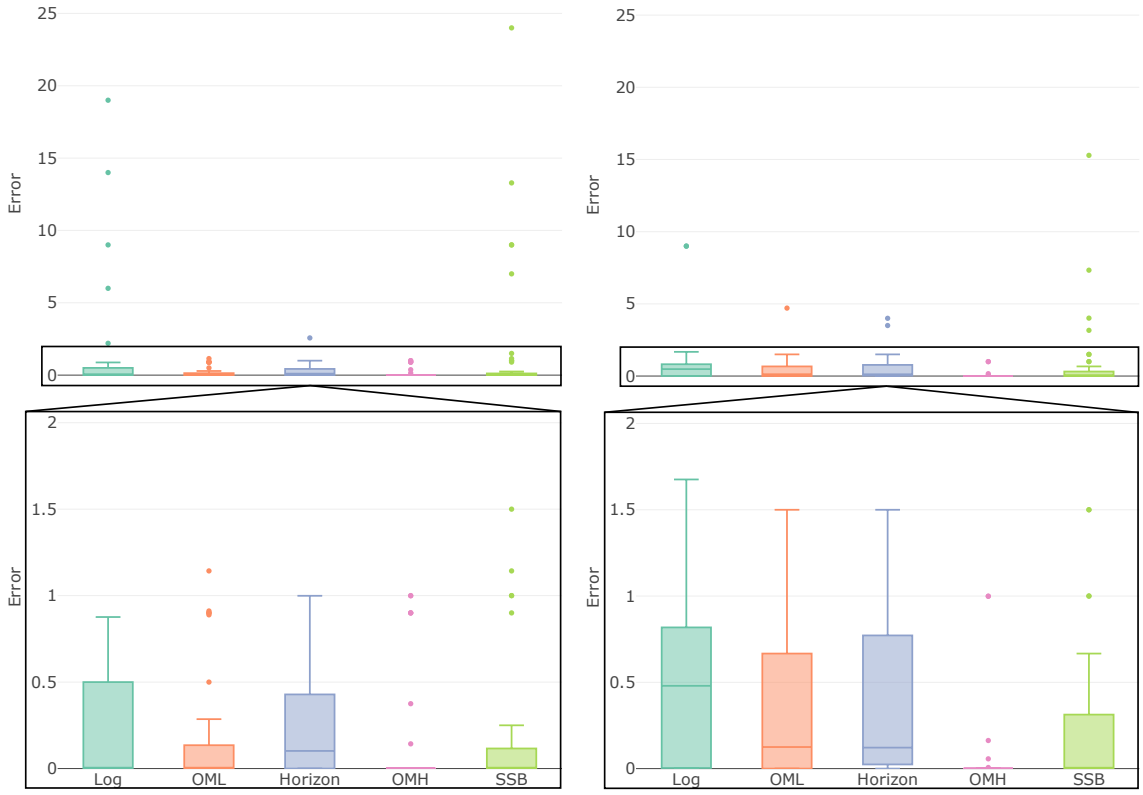
### 5.2 Error Rates

The results of the error analysis are summarized in Figure 8. For the identification and estimation tasks, the figures consist of a global box plot zoomed to a range from 0 to 2 in the lower part of the image (Figure 8a and Figure 8b). Box plots are constructed as follows: The box is bounded by the 25% and 75% quantiles of the data and contains a thicker line indicating the median. The whiskers have a maximum length of 1.5 × [box height], but only extend to the furthest data point within this range. The error bars for the discrimination and trend tasks show the mean error with the 95% confidence interval (Figure 8c and Figure 8d). In Figure 9, the error rates for the different data conditions are displayed. Averages stated below are adjusted for outliers.

**Identification Task** Our two new designs OMH and OML performed best for the identification task, while Log and SSB were most susceptible to exponent errors (see Figure 8a). The Kruskal-Wallis-test indicated a significant main effect in error rates ($\chi^2$ = 23.582, p-value = 9.686e−5). The post-hoc pairwise analysis showed (Table 1a):

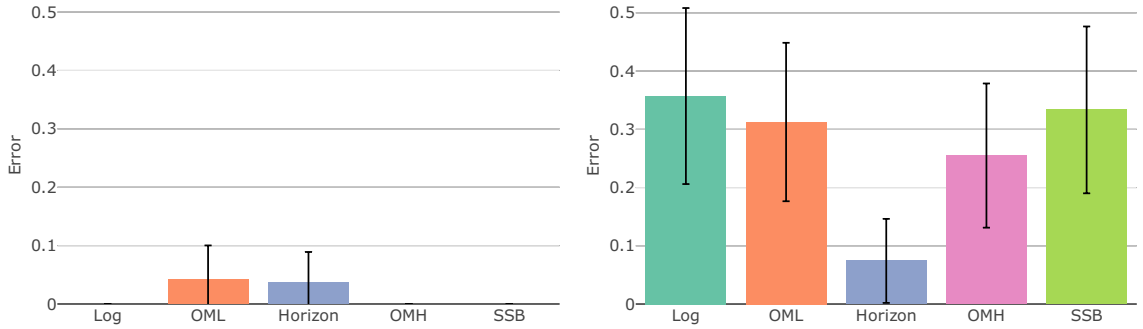
- OMH ($\bar{x}$ = 0) had a significantly lower error rate than the classic horizon graph ($\bar{x}$ = 0.21) and the log-line chart ($\bar{x}$ = 0.14).

A separate consideration of the error rates for the different data conditions in Figure 9 showed that reading values on grid lines was very easy for all designs, as almost no errors were made. For both the log-line and SSB charts, reading values at higher magnitudes resulted in more errors, while for the classic horizon graph, this was the case at lower magnitudes. Our OMH graph did not lead to increased errors in any of the conditions.




(A) Identification task (bottom: zoomed version).

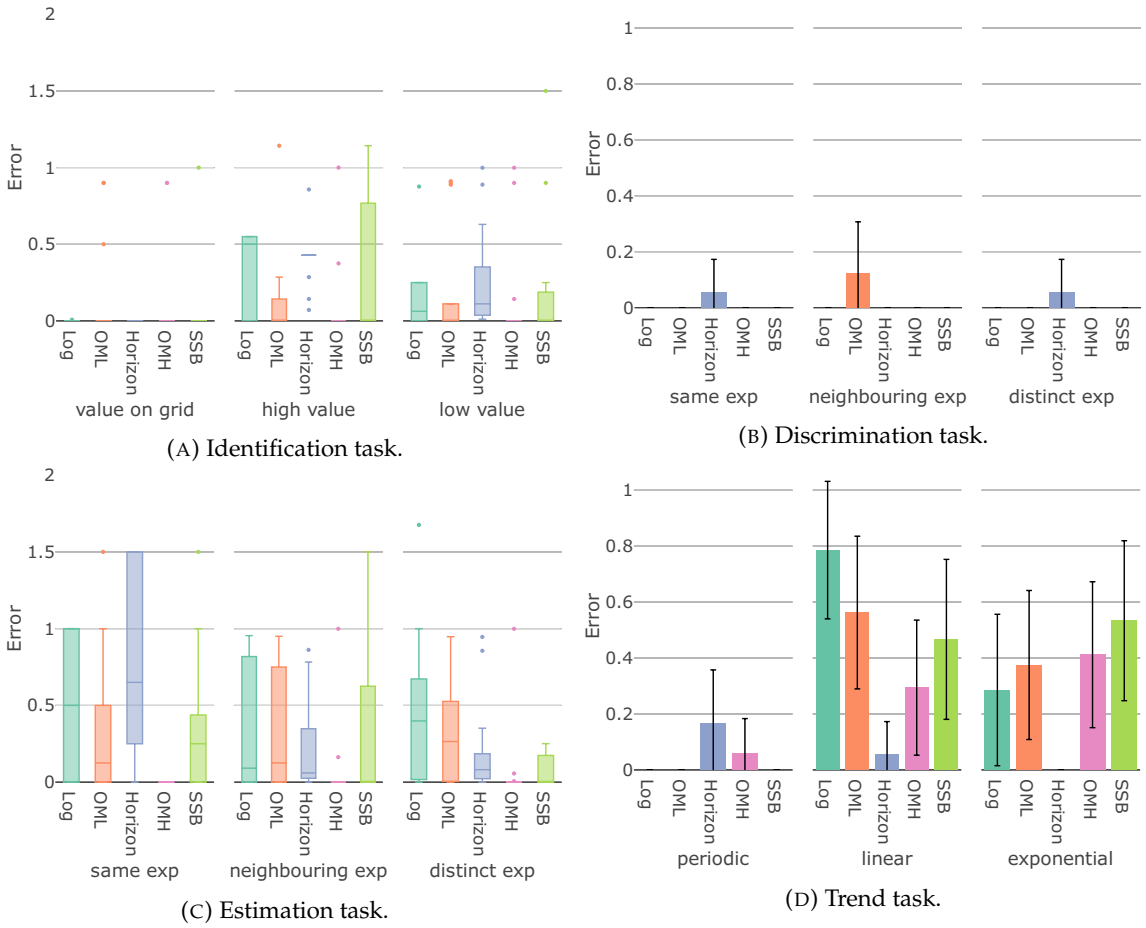
(B) Estimation task (bottom: zoomed version).




(C) Discrimination task (mean error).

(D) Trend task (mean error).

FIGURE 8: Boxplots of the errors per design for the different study tasks (the lower, the better).




(A) Identification task.

(B) Discrimination task.

(C) Estimation task.

(D) Trend task.

FIGURE 9: Error rates for the different data conditions per task and design (the lower, the better).

**Discrimination Task** In the discrimination task, only two errors each were made with the OML ($\bar{x}$ = 0.042) and classic horizon ($\bar{x}$ = 0.037) design (see Figure 8c). These very good results indicate that the participants responded to the tasks reasonably. Due to the low number of errors, no significant main effect could be determined with the Kruskal-Wallis-test ($\chi^2$ = 5.5137, p-value = 0.2385). Accordingly, no significant patterns could be observed between the different data conditions (see Figure 9).

**Estimation Task** OMH performed best in this task. SSB again had the most exponent errors, while the other three designs had similar error rates (see Figure 8b). The Kruskal-Wallis-test showed a significant main effect ($\chi^2$ = 57.564, p-value = 9.422e−12). The post-hoc pairwise analysis showed (Table 1b):

- OMH ($\bar{x}$ = 0) had a significantly lower error rate than all the other designs.
- SSB ($\bar{x}$ = 0.085) performed significantly better than the classic horizon graph ($\bar{x}$ = 0.356).

An analysis based on the data conditions showed different effects for the designs (see Figure 9). With the two line graphs – log-line and OML –, it was equally difficult to determine differences in values, regardless of the position of the values being compared. The classic horizon graph was less prone to error the further apart the two values were.

| **p-value** | *Log* | *OML* | *Horizon* | *OMH* | *SSB* |
|---|---|---|---|---|---|
| *Log* | - | | | | |
| *OML* | 1.000 | - | | | |
| *Horizon* | 1.000 | 0.054 | - | | |
| *OMH* | 0.015 | 0.582 | <0.001 | - | |
| *SSB* | 1.000 | 1.000 | 0.196 | 0.442 | - |

(A) Identification task.

| **p-value** | *Log* | *OML* | *Horizon* | *OMH* | *SSB* |
|---|---|---|---|---|---|
| *Log* | - | | | | |
| *OML* | 1.000 | - | | | |
| *Horizon* | 1.000 | 1.000 | - | | |
| *OMH* | <0.001 | <0.001 | <0.001 | - | |
| *SSB* | 0.057 | 0.611 | 0.045 | 0.004 | - |

(B) Estimation task.

| **p-value** | *Log* | *OML* | *Horizon* | *OMH* | *SSB* |
|---|---|---|---|---|---|
| *Log* | - | | | | |
| *OML* | 1.000 | - | | | |
| *Horizon* | 0.006 | 0.022 | - | | |
| *OMH* | 1.000 | 1.000 | 0.125 | - | |
| *SSB* | 1.000 | 1.000 | 0.012 | 1.000 | - |

(C) Trend task.

TABLE 1: p-values of the pairwise Wilcoxon-test for the error analysis per tasks with significant Kruskal-Wallis-test.

While most errors in SSB occurred at values in neighbouring orders of magnitude, our OMH design resulted in very low error rates regardless of the conditions.

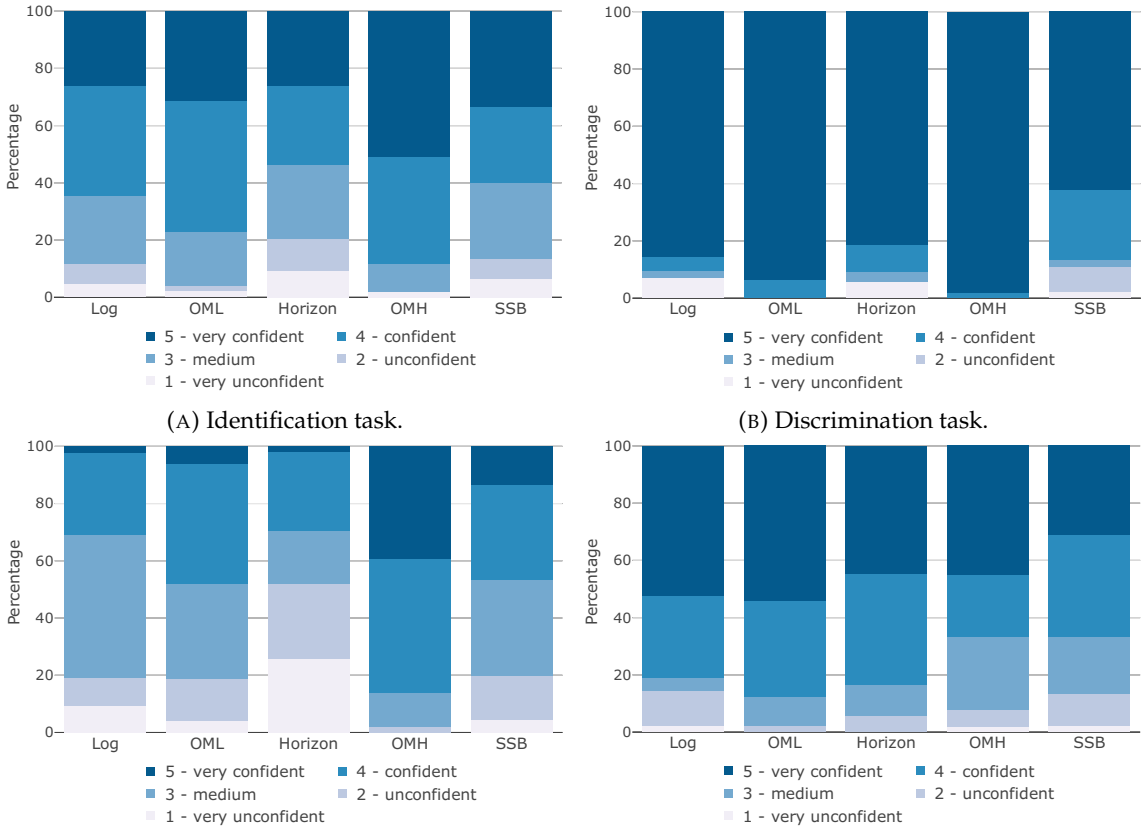
**Trend Task** For trend detection, the classic horizon graph had the lowest error rate of all designs. An obvious reason for this is the linear scaling of the y-axis exclusively in the classic horizon graph. Of the remaining designs, our new OMH and OML approaches performed best (see Figure 8d). A significant main effect in the error rates was detected by the Kruskal-Wallis-test ($\chi^2$ = 13.711, p-value = 0.008). The post-hoc pairwise analysis showed (Table 1c):

- The classic horizon graph ($\bar{x}$ = 0.074) had a significantly lower error rate than all other designs except the OMH ($\bar{x}$ = 0.255).

Figure 9 shows the data conditions. The periodic trend was very well detected by the participants in almost all designs. Note that here most of the errors occurred with the classic horizon graph, which otherwise had the lowest error rates. For the line charts OML and log-line, it was more difficult to detect the linear trend than the exponential trend, while for OMH and SSB, the opposite was true.

In sum, the error analysis shows these rankings per task:

| | |
|---|---|
| Identification: | **OMH** ≥ **OML** ≥ SSB ≥ Log ≥ Horizon |
| Discrimination: | **OMH** ~ Log ~ SSB ≥ Horizon ≥ **OML** |
| Estimation: | **OMH** > SSB ≥ **OML** ≥ Horizon ≥ Log |
| Trend: | Horizon ≥ **OMH** ≥ **OML** ≥ SSB ≥ Log |




(A) Identification task.

(B) Discrimination task.

(C) Estimation task.

(D) Trend task.

FIGURE 10: Distribution of the confidence answers per task (the higher the better).

These results partially confirm our hypothesis **H1**, although OML ranks in the middle and the standard horizon graph has the lowest error rates in trend detection.

### 5.3 Confidence

Figure 10 provides an overview of the Likert-score distributions of the confidence per task and design. The categorical responses were quantified for the means reported below, but for statistical analysis, the designs and categorical answers were tested for independence using the $\chi^2$-test. For confidence, the higher the value, the better, as this indicates the participants found the visualization easier to interpret. Hence, it provides an additional qualitative indicator of the subjective trustfulness of the visualizations in addition to the quantitative measures.

**Identification Task** The usage of our new two designs OMH ($\bar{x}$ = 4.353) and OML ($\bar{x}$ = 4.021) led to the most confidence in the identification task. The confidences for the follow-up designs Log ($\bar{x}$ = 3.738) and SSB ($\bar{x}$ = 3.733) were distributed similar, while the classic horizon graph got the least confidence ($\bar{x}$ = 3.500). Using the $\chi^2$-test, no significant dependence between designs and confidences was found ($\chi^2$ = 25.742, p-value = 0.058).

**Discrimination Task** In general, the participants were very confident in all designs for the discrimination tasks, which correlates to the very few errors made in these tasks. Our new approaches OMH ($\bar{x}$ = 4.980) and OML ($\bar{x}$ = 4.938) have even received exclusively the answers "5 – very confident" and "4 – confident" from the Likert-scale. The confidences in the log-line chart ($\bar{x}$ = 4.619) and the classic horizon graph ($\bar{x}$ = 4.611) were

| **p-value** | *Log* | *OML* | *Horizon* | *OMH* | *SSB* |
|---|---|---|---|---|---|
| *Log* | - | | | | |
| *OML* | 0.185 | - | | | |
| *Horizon* | 0.817 | 0.159 | - | | |
| *OMH* | 0.122 | 0.567 | 0.047 | - | |
| *SSB* | 0.016 | 0.006 | 0.034 | <0.001 | - |

(A) Discrimination task.

| **p-value** | *Log* | *OML* | *Horizon* | *OMH* | *SSB* |
|---|---|---|---|---|---|
| *Log* | - | | | | |
| *OML* | 0.310 | - | | | |
| *Horizon* | 0.007 | 0.007 | - | | |
| *OMH* | <0.001 | <0.001 | <0.001 | - | |
| *SSB* | 0.178 | 0.800 | 0.004 | 0.001 | - |

(B) Estimation task.

TABLE 2: p-values of the pairwise $\chi^2$-test for the confidences per tasks with significant $\chi^2$-test on all designs.

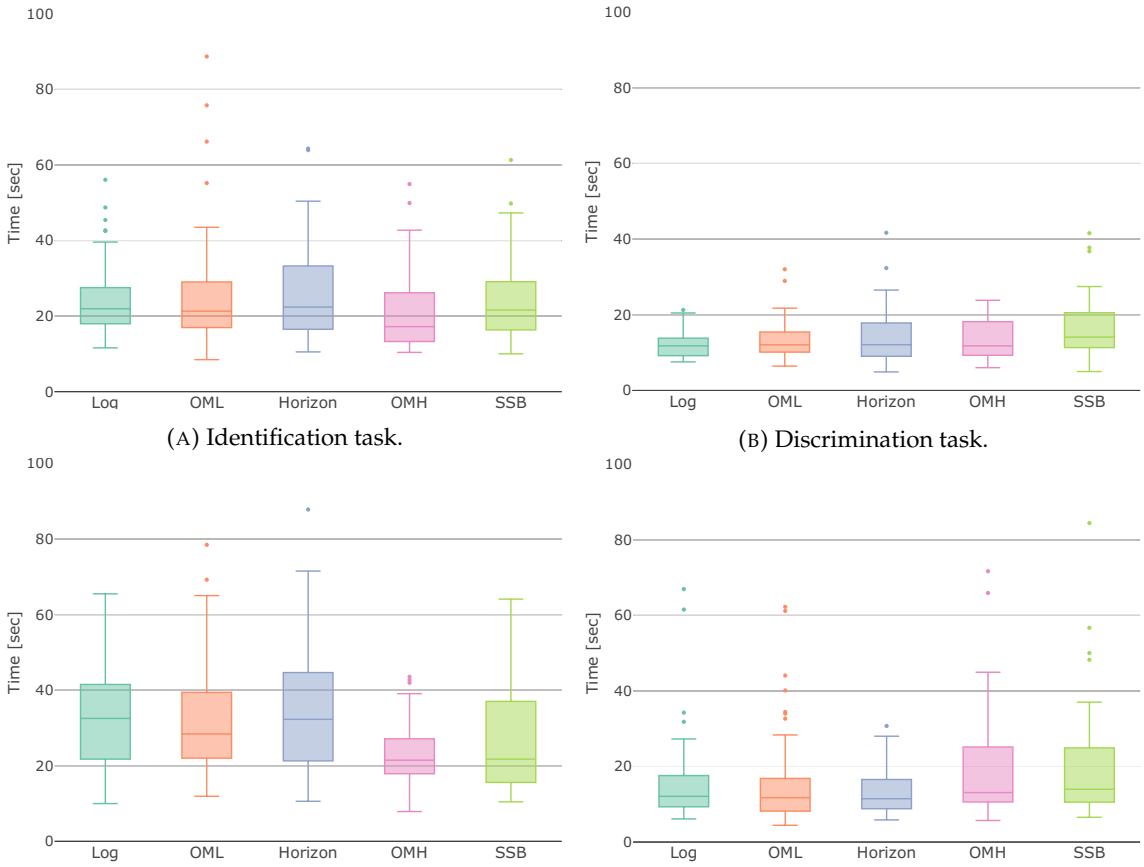
very close by, while SSB led to the least confidence ($\bar{x}$ = 4.356). The $\chi^2$-test indicated a significant dependence between design and confidence ($\chi^2$ = 47.342, p-value = 6.024e−5). The post-hoc pairwise analysis showed (Table 2a) the following significant dependencies:

- SSB led to significantly lower confidence than all other designs.
- Participants were significantly more confident using OMH than the classic horizon graph.

**Estimation Task** For the estimation tasks, our OMH graph provided the most confidence ($\bar{x}$ = 4.235), followed by the SSB chart ($\bar{x}$ = 3.356), the OML ($\bar{x}$ = 3.313), and log-line chart ($\bar{x}$ = 3.048) charts. Participants had the least confidence in the classic horizon graph ($\bar{x}$ = 2.537). A significant dependence was shown by the $\chi^2$-test ($\chi^2$ = 91.459, p-value = 1.346e−12). The post-hoc pairwise analysis showed (Table 2b) the following significant dependencies:

- OMH led to significantly higher confidence than all other designs.
- Participants were significantly less confident using the classic horizon graph than all other designs.

**Trend Task** Confidence was very close for trend detection. Despite having the lowest error rate for this task, the classic horizon graph led just to the second highest confidence ($\bar{x}$ = 4.222) behind our OML design ($\bar{x}$ = 4.396). The order behind is log-line chart ($\bar{x}$ = 4.167), OMH ($\bar{x}$ = 4.020), and SSB ($\bar{x}$ = 3.822). Due to the similar distributions, the $\chi^2$-test showed no significant dependence between designs and confidences ($\chi^2$ = 21.711, p-value = 0.153).




(A) Identification task.

(B) Discrimination task.

(C) Estimation task.

(D) Trend task.

FIGURE 11: Response times per task in seconds (the lower the better)..

Given the results from the confidence analysis, the following design rankings per task are suggested:

Identification: **OMH** ≥ **OML** ≥ Log ≥ SSB ≥ Horizon
Discrimination: **OMH** ≥ **OML** ≥ Log ≥ Horizon ➤ SSB
Estimation: **OMH** ➤ SSB ≥ **OML** ≥ Log ➤ Horizon
Trend: **OML** ≥ Horizon ≥ Log ≥ **OMH** ≥ SSB

Hypothesis **H2** is partially confirmed, since OMH and OML lead the ranking together in almost all tasks.

### 5.4 Response Time

A summary of the participants' response times in seconds is shown in Figure 11 using the same box plot design as for the error rates, with lower values being better.

**Identification Task** Participants responded fastest with our OMH design for the identification tasks ($\bar{x}$ = 21.228). The Log ($\bar{x}$ = 24.470) and SSB ($\bar{x}$ = 24.641) had almost equal response times being three seconds slower, while the classic horizon graph ($\bar{x}$ = 26.183) and OML ($\bar{x}$ = 29.689) were the slowest. The Kruskal-Wallis-test showed no significant main effect ($\chi^2$ = 7.874, p-value = 0.096).

**Discrimination Task** The discrimination tasks were overall the tasks with the lowest response times, and all designs have a maximum mean difference of less than five seconds. The log-line chart was the fastest ($\bar{x}$ = 12.142), followed by OML ($\bar{x}$ = 13.299), OMH

| **p-value** | *Log* | *OML* | *Horizon* | *OMH* | *SSB* |
|---|---|---|---|---|---|
| *Log* | - | | | | |
| *OML* | 1.000 | - | | | |
| *Horizon* | 1.000 | 1.000 | - | | |
| *OMH* | 1.000 | 1.000 | 1.000 | - | |
| *SSB* | 0.019 | 0.207 | 0.578 | 0.242 | - |

(A) Discrimination task.

| **p-value** | *Log* | *OML* | *Horizon* | *OMH* | *SSB* |
|---|---|---|---|---|---|
| *Log* | - | | | | |
| *OML* | 1.000 | - | | | |
| *Horizon* | 1.000 | 1.000 | - | | |
| *OMH* | <0.001 | 0.004 | 0.004 | - | |
| *SSB* | 0.098 | 0.192 | 0.196 | 1.000 | - |

(B) Estimation task.

TABLE 3: p-values of the pairwise Wilcoxon-test for the response time analysis per tasks with significant Kruskal-Wallis-test.

($\bar{x}$ = 13.414), and the classic horizon graph ($\bar{x}$ = 14.131), while the participants needed the most time for the SSB chart ($\bar{x}$ = 16.853). A significant main effect in the error rates was detected by the Kruskal-Wallis-test ($\chi^2$ = 9.962, p-value = 0.041). The post-hoc pairwise analysis showed (Table 3a) only one significant difference:

- The log-line chart had significantly lower response times than the SSB design.

**Estimation Task** The estimation tasks took the longest of all four tasks, and the response times varied the most between designs, with a maximum mean difference of more than 10 seconds. Our OMH method was the fastest ($\bar{x}$ = 22.964), which is consistent with the error results, as were the second lowest response times of SSB ($\bar{x}$ = 26.249). Log ($\bar{x}$ = 32.819), OML ($\bar{x}$ = 33.666), and Horizon ($\bar{x}$ = 34.094) were the order behind it. Due to the high variation in the response times, a significant main effect was indicated by the Kruskal-Wallis-test ($\chi^2$ = 23.834, p-value = 8.625e−5). The post-hoc pairwise analysis showed (Table 3b) only the following significant differences:

- Our OMH design was significantly faster than all other designs except SSB.

**Trend Task** Matching the error and confidence results, the classic horizon graph got the fastest responses ($\bar{x}$ = 13.296). This was followed by the two line graphs Log ($\bar{x}$ = 15.751) and OML ($\bar{x}$ = 16.348), while the OMH ($\bar{x}$ = 19.396) and SSB ($\bar{x}$ = 20.520) tasks were the slowest to complete. The Kruskal-Wallis-test showed no significant main effect ($\chi^2$ = 8.651, p-value = 0.070).

The following design rankings per task are suggested based on the response time analysis:

| | |
|---|---|
| Identification: | **OMH** ≥ Log ≥ SSB ≥ Horizon ≥ **OML** |
| Discrimination: | Log ≥ **OML** ≥ **OMH** ≥ Horizon ≥ SSB |
| Estimation: | **OMH** ≥ SSB ≥ Log ≥ **OML** ≥ Horizon |
| Trend: | Horizon ≥ Log ≥ **OML** ≥ **OMH** ≥ SSB |

With these results, hypothesis **H3** is rejected, since in almost all cases, OMH and OML are not significantly worse than the others. In fact, there are among the fastest for some tasks.

### 5.5 Free Text Feedback

We received open-ended feedback from 30 participants at the end of the study. From this feedback, the introductions to the tasks and visualization techniques seemed to be understandable for the participants.

We got most (8) comments on the OMH graph. Participants found our approach to be exciting and mentioned that "using color to present the exponential is a very good visualization method". It was noted that the choice of colors could be improved to make it easier to distinguish between the bands. Using color as a visual channel to make it easier to read accurate values was also mentioned for the OML chart. In general, the responses to both the OMH and OML designs were quite positive. Moreover, the feedback confirmed the results that "trend detection was [the] most difficult" task with the OMH design.

The standard horizon graph received mixed feedback. While it was noted that the design was helpful for discrimination, it was criticized that the overlapping areas with the corresponding value ranges distracted from the actual content of the graph.

For the log-line graph and the SSB chart participants indicated that these designs were "not suitable for reading concrete data or determining the difference".

## 6. Discussion

Order of magnitude horizon (OMH) graphs either outperform significantly or perform comparably well to their currently used counterparts (log-line charts, SSB, and classic horizon graphs) in identification, estimation, and discrimination tasks in terms of error, confidence, and response times. Order of magnitude line (OML) charts, however, do not perform statistically better on error and response times. This confirms the conclusions of previous works [23,29,93,94] from the area of visualizing large value ranges, that the separate representation of mantissa and exponent is suitable for large value ranges (**H1**). Moreover, it appears that this separation needs to be *discrete*, as the continuous separation in the case of OML does not seem to be sufficient. For trend detection, the classic horizon graph provides significant lower error rates compared to all designs but OMH (lower, but not significant). Horizon graph's significantly better performance is an indication that linear axis scaling is more appropriate for this type of task. It is interesting to note the comparable performance of the two horizon-based designs, with OMH exhibiting signs of speed-accuracy trade-off (SAT) effects (even if not significant) which may indicate participants trading speed for accuracy while undertaking a complex task with a new design. An interesting line of further investigation would be to examine if increased familiarity and usage of OMH could potentially reduce this apparent trade-off and increase performance similarity. These could indicate an inherent effect of the design itself which leverages perceptual grouping.

In general, however, the results of the individual designs seem to be dependent on the values of the queried data points and their ratios, as the analysis of the different data conditions shows. For example, the SSB seems to be particularly difficult to read in high orders of magnitude, despite its design being intended to favor large value ranges. Furthermore, the exponential trend is easier to detect than the linear trend in the two line charts (log-line and OML), whereas the opposite is true for OMH and SSB. In summary, further guidelines for using different designs in specific data and task conditions may emerge.

Another interesting aspect is the advantage of using color as a visual variable. The large amount of errors in the exponent of the log-line and SSB charts in identification and estimation tasks suggests that color coding supports the detection of the correct exponents, as these are the only two designs not using color as a visual variable. An additional indicator for this is that the participants had the highest confidence with either OMH or OML (which use color coding) in all tasks (**H2**). Since OMH and the classic horizon graph both use the design principles of the two-tone pseudo coloring [160], the improved performance of OMH is most likely due to the novel y-axis composition.

Our OML technique uses the order of magnitude colors (OMC) colormap of Braun et al. [29], and our results support the usefulness of their approach as it leads to improved error rates compared to the standard log-line chart. The OMC color favors "banding effects" when crossing magnitude thresholds, which seems to have contributed to improve the overall visual clarity in our OML layout, although we have not formally tested this in our study. The improved error rates suggests OML to be worthy of further investigation in those domains where not only large value ranges are present, but quantitative discrimination of these magnitudes is a task of major importance. In addition, as pointed out by Borkin et al. [24], the low data-to-ink ratio in OMH and OML is an indication that these two visualizations may be more memorable than standard visualization techniques.

Participants had no previous experience with our new visual layouts, and we carefully designed our studies to minimize possible learning effects. We were therefore positively surprised by the fact that no significant global trade-off effect in the response times, confidence, and error rates was detected. Only (not significant) SAT symptoms were detected for OML in the Identification task, and OMH in the Trend detection tasks (as previously discussed), with participant trading speed for accuracy in both cases. This seems to suggest that the cognitive load introduced by the new design was negligible (**H3**). The consistency of the subjective confidence responses with the performance of the designs implies that our new approaches were correctly interpreted. It would be interesting to explore further how easily our novel design can be learned when applied to more complex scenarios, as also suggested in [202].

### 7. Limitations and Future Work

There are still some limitations to our visualization techniques that subsequently provide opportunities for future work.

As the error rates analysis showed and the participant's feedback confirmed, our designs are not well suited for trend detection. It would be interesting to see if an approach can be developed that works well for trend tasks and one or more of the other tasks. Despite the great results of our OMH approach, participants' comments also indicate that there is room for improvement in the choice of colors for the different orders of magnitude. The analysis also showed slower response times of OML charts compared to log-line charts. This could be either due to the novelty of OML charts, or the increased cognitive load from double encoding the values by position and color. Although the free feedback suggests the former, this cannot be proven by our study and requires further investigation.

Although a participant pool from university is not representative for the entire society, it made sure that everyone knew the exponential notation. We are aware that our study covers only part of the possible types of tasks and data, and that it would be interesting to investigate different types of both (e.g. multiple time-series). Nevertheless, it was sufficient to test our hypotheses. In particular, value scales with both positive and negative values have not been considered in this paper. As mentioned in Subsection 3.1,

this can be relatively easily implemented for the OMH design. For the OML chart, on the other hand, further considerations are necessary.

Our approach leverages the effectiveness of visual encoding rather than relying on user interaction (e.g., zooming). The ability to visualize the full data range reduces the need for direct manipulation of the visualization. This has the advantage of being able to maintain context and ensure consistency across different analytical tasks, as well as being more suitable for print. However, it may be possible to integrate and adapt interaction techniques to further increase the effectiveness (especially for long time-series) of the visual encoding in future work.

Finally, our work has focused on orders of magnitude with base 10, as these are the most common. Base 2 and base $e$ would require more colors due to the high exponent range, and vice versa for larger bases.

## 8. Conclusion

Our work presents two novel visualization designs to support the display of large value ranges in time-series data: The order of magnitude horizon chart and the order of magnitude line chart. Using an empirical user study, we showed the advantages of our designs in accuracy, confidence, and response time for identification, discrimination, and estimation tasks. The performance of our visualization designs for large value ranges is partially dependent on the order of magnitude of the values to be evaluated, with values between magnitudes being harder to compare. Our results confirm previous works in the field of visualization of large value ranges [23, 29, 93, 94] that the separation of the mantissa and exponent is beneficial for the perception of values from large value ranges.

## Acknowledgments

The authors would like to thank all study participants for taking part in the evaluation and the reviewers, whose suggestions helped to improve this paper. This work has been partially supported by BMBF WarmWorld Project and KPA Intelligent Methods for Earth System Sciences.

## 2.3 Design and Evaluation of Visualizations for Large Value Ranges in Multiple Time-Series

The latest work on the visual communication of large value ranges combines two challenging problems by visualizing large value ranges in multiple time-series. The design space for large value ranges visualizations in time-series data is explored. Subsequently, seven designs (standard approaches, the techniques from the individual time-series research plus adaptions, and one novel design) are evaluated in two user studies using data containing large value ranges and common tasks of multiple time-series research.

The paper is accepted for publication in Information Visualization journal:

**D. Braun**, R. Borgo, M. Sondag, and T. von Landesberger. Design and Evaluation of Visualizations for Large Value Ranges in Multiple Time-Series. *Information Visualization*, 2025. doi: 10.1177/14738716251349501

The supplementary material of the paper, including the study preregistration, data, results, and documentations as well as the Python code for the data and design generation, is publicly available at OSF.

I am the primary author of this publication. In this role, I was responsible for the design, implementation, data collection and analysis, as well as the writing and publication of the work. The specific contributions of myself and my co-authors to this publication are outlined below according to the Contributor Roles Taxonomy (CRediT):

**D. Braun**: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Software, Visualization, Writing – Original Draft. *M. Sondag*: Conceptualization, Methodology, Investigation, Visualization, Writing – review & editing. *R. Borgo, T. von Landesberger*: Supervision, Conceptualization, Methodology, Writing – review & editing.

# Design and Evaluation of Visualizations for Large Value Ranges in Multiple Time-Series

DANIEL BRAUN[1], RITA BORGO[2], MAX SONDAG[1], TATIANA VON LANDESBERGER[1]

[1]*University of Cologne*
[2]*King's College London*

**Abstract:**

This paper investigates the complex issue of visualizing large value ranges in multiple time-series. We propose the design spaces for this composed visualization. In two crowdsourced user studies, we test seven designs: Three state-of-the-art designs, three extensions to existing designs, and one novel design. We assess five tasks: Maximum and minimum identification, value discrimination, difference estimation, and slope assessment. Our results show novel findings: For the minimum task, where values in low orders of magnitude have to be identified, our novel height-stack line chart yields the best results. For slope assessment and all tasks where the maximum value can be used as a proxy for the correct answer (maximum, discrimination, and estimation), the linear line graph shows comparable results to all other designs. Moreover, the use of visual mapping to color supports the perception of mantissa and magnitude variations. Unexpectedly, our results indicate that increasing the number of time-series does not generally reduce the accuracy of estimation, discrimination, and identification. Our findings are domain independent. They provide useful insights for designers seeking to visualize large value ranges in multiple time-series, e.g., for financial, medical, or meteorological data.

*Information Visualization, 2025*

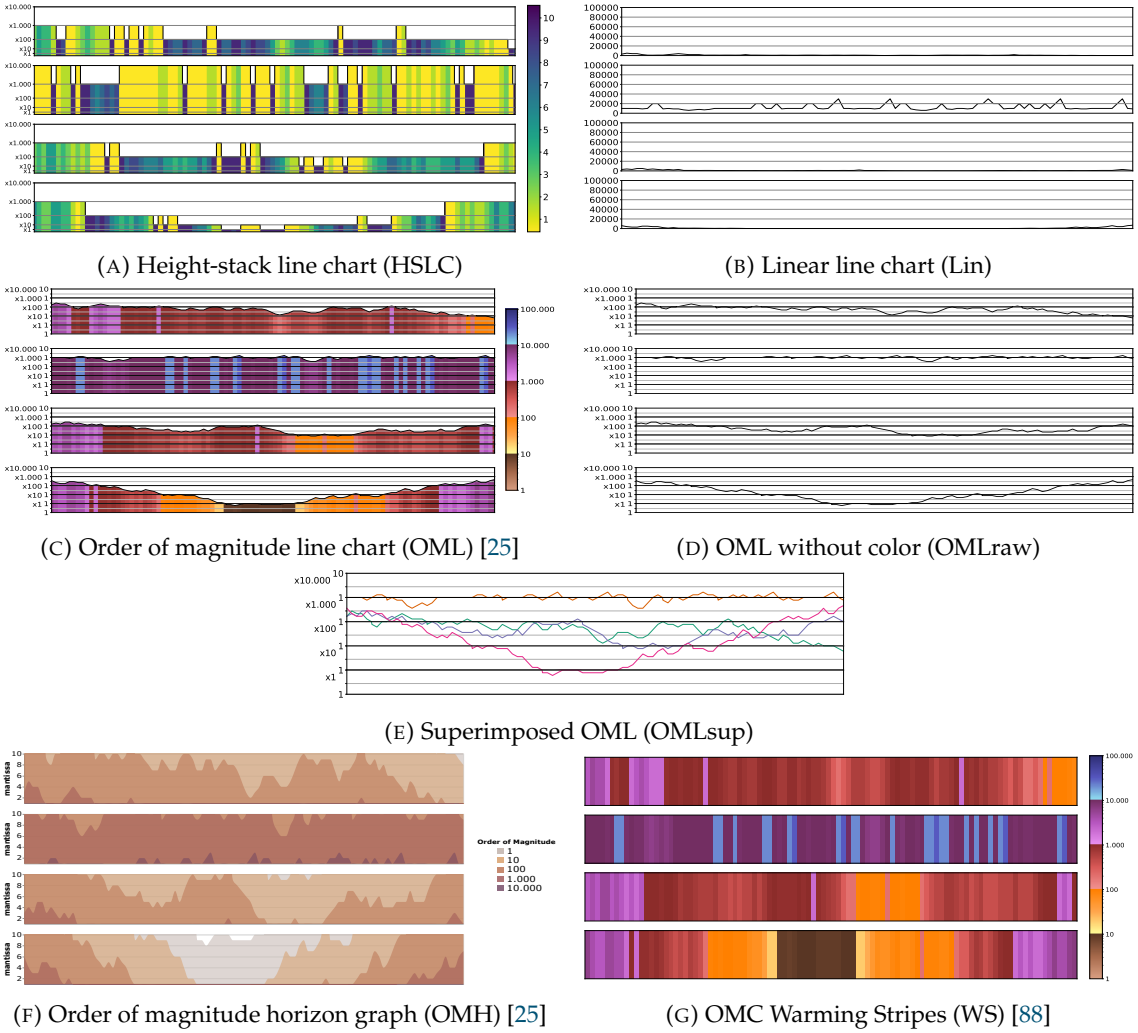
## 1. Introduction

Time-varying data featuring large value ranges (i.e., time-series with range of values of several orders of magnitude) are increasingly common across various domains including finance, medicine and beyond. A familiar example is the comparison of COVID-19 cases across different regions which can easily exhibit ranges from zero to one hundred thousand [121].

Standard time-series visualization designs – such as linear line charts – work well for small value ranges [4, 68]. However, they fall short in effectively displaying large orders of magnitude, with small values "hidden" by the large range of the axis. Solutions for showing both small and large values have been previously investigated for individual data sets [23,25,93,94]. For example, Braun et al. [25] successfully developed new designs for showing large values ranges within a single time-series, although open questions remained on the extent of the contribution of using color to the readability.

The complexity of displaying large value ranges is further increased when considering not just a single time-series, but multiple time-series. This changes the tasks to be performed on the visualizations, since comparing is always required. Furthermore, the screen space available for the individual visualizations decreases as the number of time-series increases, making it more difficult to identify details.

Hence, it is unclear whether the efficacy observed in single time-series extends to comparisons involving multiple time-series. Moreover, existing literature on multiple time-series focuses exclusively on data sets with small value ranges [75,76,100,167]. Thus, we aim to resolve the following research questions:



(A) Height-stack line chart (HSLC)

(B) Linear line chart (Lin)

(C) Order of magnitude line chart (OML) [25]

(D) OML without color (OMLraw)

(E) Superimposed OML (OMLsup)

(F) Order of magnitude horizon graph (OMH) [25]

(G) OMC Warming Stripes (WS) [88]

FIGURE 1: The evaluated visualization designs in our user study all showing the same example tested data set: our new design *height-stack line chart* (a), the standard linear line chart (b), the state of the art designs for large value ranges in single time-series – the order of magnitude line chart [25] (c) with the two variants without color (d) and superimposed (e) and the order of magnitude horizon graph [25] (f) –, the OMC warming stripes [88] (g).

- **RQ1**: Are state-of-the-art visualization designs for large value ranges in single time-series also effective for visualizing multiple time-series?
- **RQ2**: To what extent does the visual mapping to color contribute to the readability of magnitude variations in large value range visualizations?
- **RQ3**: How does the performance of our novel height-stack line chart compare to existing visualization designs for multiple time-series with large value ranges?

In order to find suitable combinations of visualization designs for both areas, we explore the design spaces for large value ranges and multiple time-series visualizations including both juxtaposed and superposed layouts.

As a result, we compared seven visualization designs: Three of which represent state-of-the-art designs for each of the two domains (the standard linear line chart and the two designs from Braun et al. [25]), three of which are extensions of existing designs (adaptions of the order of magnitude line chart and warming stripes), and the newly designed height-stack line chart (see Figure 1).

We conducted two crowdsourced user studies with 105 participants each to answer the research questions.

In the main study, we evaluated all designs on four low-level tasks [10,31,187] that are common for using visualizations of large value ranges [23,25,29,93,94] and multiple time-series [75, 76, 100, 141, 190]: maximum identification and reading, value discrimination, difference estimation and slope assessment.

For all tasks, we test different levels of difficulty with respect to the number of time-series (2 and 4) and data properties (magnitude differences between the series).

All designs perform comparably in terms of accuracy to the baseline of the linear line chart for the tasks tested, while being significantly worse in terms of task completion time. Surprisingly, in contrast to research on the accuracy of small value ranges [100], the accuracy of the adapted designs for large value ranges seems to *increase* in several cases when more time-series are presented. Moreover, the use of color in visualizing large value ranges supports their readability.

Based on our findings regarding the previous four tasks, we conducted a follow-up study containing the minimum task to test the special properties of large value ranges that the entire value range (including small values, which often have low discriminability) is important. For minimum identification, our height-stack line chart (HSLC) design outperforms all other techniques. This indicates that our new design works well across the entire value range.

The remainder of the paper is structured as follows: We first present an overview of current methods and evaluations (Section 2). Then, we present the design spaces and the visualization designs tested in the studies in Section 3. Afterward, the structure of our user studies is described (Section 4), followed by an exposition of their analysis and results (Section 5). At the end, we interpret our findings (Section 6) and discuss limitations and open research questions (Section 7).

## 2. Related Work

There are three main related areas of research: Visual comparison in general, visualizations for multiple time-series in specific, and visualizations for large value ranges.

**Visual Comparison** Gleicher et al. [78,79] provided a taxonomy of comparative designs that groups designs in three categories: *Juxtaposition* designs present each object separately (e.g., small multiples), *superposition* designs present multiple objects overlaying in the same coordinate system, and *explicit encoding* designs directly visualize connections between objects. In 2021, L'Yi et al. [118] used this taxonomy as a basis to survey the current state-of-the-art in comparative layouts, and summarized their results as a set of guidelines for visualization designs. We use these taxonomies and guidelines as a basis for our design decisions for our experimental setup and visualization choice.

**Visualization of Multiple Time-Series** While there is much research on visualizing single time-series [2,4,6–8,33,34,51,68,81,90,98], we focus on multiple time-series.

Javed et al. [100] compared juxtaposed time-series visualizations with superimposed visualizations for identification, discrimination, and slope tasks. They found that split-space (i.e., juxtaposed) techniques were more efficient for comparisons with a large visual span, while shared-space (i.e., superimposed) techniques were more efficient for smaller visual spans. These results are in contrast to the results of other studies [118, 135], in which superimposed layouts surpassed others in all tasks. Franke et al. [75] confirmed that the performance of visualization techniques for multiple time-series is task dependent. This large variance in results for different tasks prompted us to test both superimposed and juxtaposed visualizations in our study.

Many [76,89,99,110,141] visualization designs have been proposed and evaluated for multiple time-series. We will consider this variety of designs to determine whether they are extendable to supporting large value ranges.

**Visualization of Large Value Ranges** Hlawatsch et al. [93] were the first showing that visualizations designed especially for the characteristics of large value ranges outperform standard techniques in accuracy and response time. A common [23,25,29,94] approach in visualizing large value ranges is the separate representation of mantissa and exponent. It is, however, not yet clear which mapping of the mantissa and exponent to which visual variables allows for the best results.

For time-dependent data, line charts with logarithmic scaling are the most common visualization technique for time dependent data with large value ranges [111, 155, 164]. However, other approaches do exist, such as the order of magnitude horizon graph [25] (Figure 1f) and order of magnitude line chart [25](Figure 1c), which make use of color to visualize either the exponent $e$ or mantissa $m$ of a value $v = m \cdot 10^e$. These designs outperformed the standard log-line charts for single time-series. We will include these designs in our study to evaluate whether these results extend for multiple time-series.

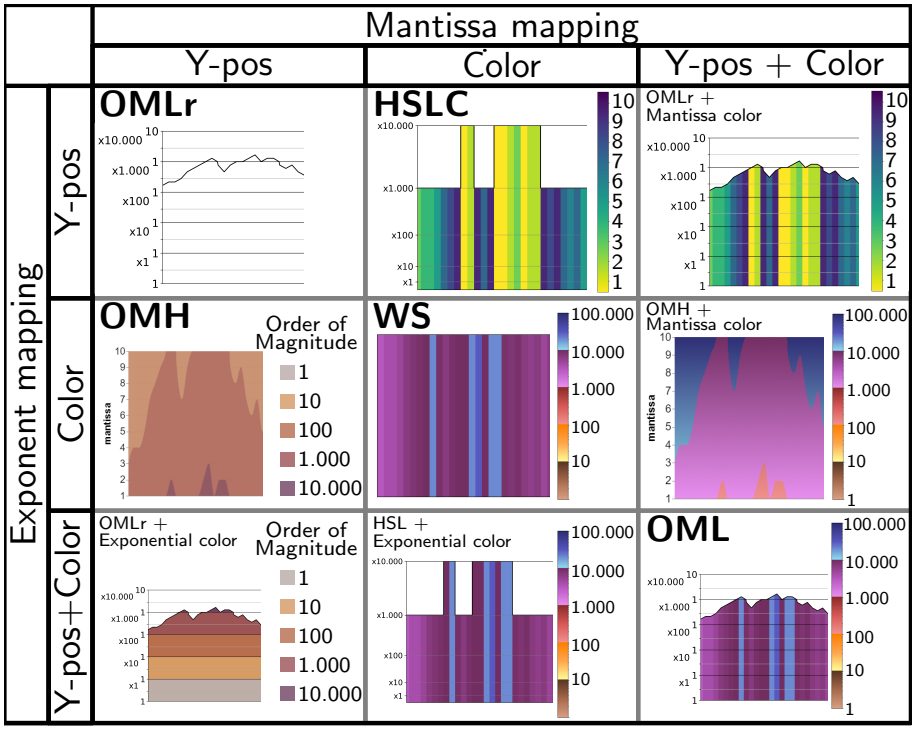



FIGURE 2: Design space for large value ranges in time-series data: examples of visualization designs for mapping mantissa and exponent to Y-position and/or color.

## 3. Visualization Designs

The combination of large value ranges and multiple time-series adds complexity to the design process. With multiple series, the screen space available for the individual visualizations decreases with the number of time-series. Moreover, each task includes a comparative component. To be comparable to previous research [4, 25, 29, 75, 100], we restrict ourselves to non-interactive visualizations that can also be used in printed media. In order to determine which design compositions are promising, we first explore the design spaces of both areas. Afterwards, we describe the individual designs tested in the user study.

### 3.1 Design Space - Potential Designs

**Design Space for Large Value Ranges** Current designs [23, 25, 29, 94] for large value ranges in time-series, all map the time to the x-position. They then split a value $v = m \cdot 10^e$ into a continuous mantissa $m \in [1, 10)$ and a discrete exponent $e$, which are mapped to the y-position and/or color [18]. This separation allows for a direct comparison of mantissa and exponents of two or more values or time-series.

The design space for showing individual time-series is spanned by the combinations of visual mappings to y-position and to color for mantissa and exponent. We enumerate the combinations in Figure 2 and show examples of visualization designs for each combination. Additionally, we fill in the unexplored combinations through changing the color scheme of existing visualization designs. For the mapping of mantissa to color and exponent to y-position, simply changing the color scale was not sufficient. Thus, we created a new design: The height-stack line chart (HSLC).

**Design Space for Multiple Time-Series** Besides the existing techniques for visualizing multiple time-series [76,89,99,100,110,141], one additional dimension of the design space for showing multiple time-series is whether juxtaposition, superposition, or explicit encoding is used [78, 79]. As explicit encoding would require a complete redesign of the visualization, and hence is out of the scope of this paper.

The designs in Figure 2 can all be represented using juxtaposition easily. Using superposition is however more complicated. The standard way of identifying different series in superposition is mapping each series to a color. However, all but one of the designs (OMLr) identified, already have a mapping to color. For OMH and HSLC it would technically be feasible to split the color variable in hue and lightness, but this would result in an unpractical amount of overlap. An alternative braided design as considered by Javed et al. [100] was also not recommended for multiple time-series. Hence, we only include a superimposed variant of OMLr as a superimposed design, resulting in OMLs.

### 3.2 Tested Designs

Testing all combinations of the designs space would have overwhelmed the evaluation (see Figure 2). To prioritize the most promising designs from previous studies [25], we exclude mappings that only require color changes, as we do not expect these to perform significantly different. Thus, all bolded designs in Figure 2 (OML, OMLr, OMH, HSLC, WS) were tested. Moreover, we tested a juxtaposed linear line chart (Lin) and a superimposed version of the OMLr chart (OMLs) as baseline designs (see Figure 1).

**Height-Stack Line Chart (HSLC)** Our novel design – the so-called "height-stack line chart" (Figure 1a) – builds upon techniques of large value ranges by separately representing the mantissa and exponent. We map the mantissa to color, and the exponent to the (discrete) y-position. The novel representation of the exponent on the y-position supports the perception of magnitude changes, especially in the context of reduced pixel availability with multiple time-series. For the color scale, we chose to use viridis. As a perceptually uniform, multi-hue color scale, it allows for the best possible perception of the mantissa in the small range of $[1-10)$ [105, 134, 200] and preserves the "dark is more" bias [83]. Even though the data in our study assume discrete mantissas, we use a continuous color scheme to be able to map the entire possible value range.

To enhance the perception of absolute value changes, we adjusted the scaling of HSLC's y-axis to be non-linear. This has the advantage that the difference in heights show relative value changes. The distances between the discrete exponents y-axis tick marks increase quadratic (i.e., a value $v = m \cdot 10^e$ is drawn at $y = e^2 + e + 1$) to strike a balance between visibility and screen space. Although we consider only positive values in our study, HSLC can easily be extended for negative values with a divergent color scale.

**OMC Warming Stripes (WS)** The original "warming stripes" (WS) are charts developed to show climate changes over time [88, 170]. Nowadays, however, they have been used for other applications with time-dependent data as well. Warming stripes consist of narrow vertical segments (stripes) arranged horizontally from left to right. Each stripe represents one time step. The arrangement from left to right corresponds to the time on the horizontal axis. The value of each time point is indicated by the color-coding of the respective stripe. This can save a lot of vertical screen space by mapping data values to color instead of y-position, as in line or bar chart.

We adapt this approach to the challenges of large value ranges – so-called "OMC WS". We combine warming stripes with the order of magnitude color scheme (OMC) developed by Braun et al. [29]. The OMC scale is specifically designed for large value ranges

by using a different hue for each exponent and a sequential scale of the respective hue for the mantissa. The change in hue between the different orders of magnitude reduces the probability of magnitude errors. An example of the result can be seen in Figure 1g.

**Order of Magnitude Horizon Graph (OMH)** The Order of magnitude horizon graph (OMH) (Figure 1f) is a design developed for showing single time-series with large value ranges. In the study by Braun et al. [25], this design outperformed all other tested designs. OMH is an adaptation of the standard horizon graph [71,150,160]. Horizon graphs divide the value range into multiple bands of equal size. The bands are then overlaid on a shared y-axis and color is used to distinguish the individual bands.

In the OMH design, each band represents a different order of magnitude, with the number of bands depending on the number of orders of magnitude. The coloring of the bands is based on the hue gradient of the OMC color scale: The larger the magnitude of a band, the higher the saturation. By construction, the mantissa is mapped to a continuous y-axis. The resulting visual mapping swaps the visual variables of exponent and mantissa compared to our HSLC design.

**Order of Magnitude Line Chart (OML)** The second design introduced by Braun et al. [25] for single time-series with large value ranges – the order of magnitude line chart – is an extension of the logarithmic line chart. It uses linear scaling within each order of magnitude for easier reading of the mantissa and colors the area below the line in the graph using the OMC color scale (Figure 1c). Thus, each value is double-encoded by the visual variables y-position and color.

**OMLr:** An open question that arose from the work of Braun et al. [25] was whether the reason for the improved results of the design was the improved axis scaling or the use of color (our RQ2). To answer this question, we also include the *OML design without color (OMLr)* (Figure 1d).

**OMLs:** We also consider the superimposed variant of OML. Here, we modify OMLr to a superimposed chart, where each different time-series is colored using a categorical color scale from ColorBrewer [86] (Figure 1e).

**Linear Line Chart** Finally, we include the linear line chart as a baseline to compare against. The logarithmic line chart was already compared in previous work [25], and hence we do not include it in this study.

### 4. Evaluation

We performed two users studies to evaluate the seven visualization design, including state-of-the-art designs (Lin, OMH, OML), extensions of existing designs (OMLr, OMLs, WS), and one novel design (HSLC). In the first study we compare these seven designs in four different tasks (maximum identification, value discrimination, difference estimation, and slope assessment). Our experiment hypotheses and test plan were preregistered on OSF. We have changed the presentation of these hypotheses for increased clarity of reading. The content and statistical tests performed remain the same.

- **H1, H2**: *Our new HSLC design reduces error rates (H1) and response times (H2) in maximum, discrimination, estimation, and slope tasks on large value ranges in multiple time-series compared to state-of-the-art designs for large value ranges in single time-series.*
  We expect that mapping the exponent to the y-position reduces exponent errors due to an increased perception of magnitude variations. Thus, we expect participants to be more accurate and faster with HSLC.

- **H3, H4**: *The order of magnitude horizon graph outperforms all variants of the order of magnitude line chart in maximum, discrimination, estimation, and slope tasks in terms of accuracy (H3) and response time (H4).*
  We expect that the results of Braun et al. [25] for large value ranges in single time-series (i.e., OMH outperforming OML) will also apply to large value ranges for multiple time-series.

- **H5, H6**: *Maximum, discrimination, estimation, and slope tasks in order of magnitude line charts are more accurate (H5) and faster (H6) with double-encoding of color and position.*
  We expect the double-encoding via color of the OML design to improve readability of exponent variations. With double-encoding, viewers could use their preferred visual variable to determine the answer and the second variable to validate it.

- **H7, H8**: *The linear line chart is less accurate (H7) and slower (H8) than designs for large value ranges in maximum, discrimination, estimation, and slope tasks.*
  Linear line charts work well for multiple time-series with small value ranges. However, visualizations explicitly designed for large value ranges should enhance the readability for this type of data, as previous studies [23, 25, 29, 93, 94] proved for other data types.

In the second follow-up study (performed after seeing the results of the first study) we additionally tested the minimum identification task to gain more insight into how the designs perform on the entire value range in multiple time-series. In all other regards, the second study follows the same structure and procedure as the first study. We did not preregister hypotheses for this second study, but expected the outcome of the hypotheses **H1-H8** of the first study to apply equally to the second study.

### 4.1 Tasks

We evaluate five low-level tasks that follow state-of-the-art task taxonomies [10, 31, 187]: Maximum and minimum identification, value discrimination, difference estimation and slope assessment. These tasks are composed of typical tasks from the two research areas related to our work: identification, discrimination and estimation are standard tasks in large value ranges research [23, 25, 29, 93, 94], while maximum, discrimination, and slope tasks are commonly used in studies on multiple time-series [75, 76, 100, 141]. The minimum task (tested in the second study) additionally aims to test whether the designs can appropriately display lower orders of magnitude in large value ranges. One major difference between single and multiple time-series tasks is that each task inevitably contains a comparison component with more than one series.

For each of the five tasks, we will consider combinations of two different **difficulty levels**: The amount of time-series, and the exponent difference between the time-series.

For the amount of time-series, we will present either 2 or 4 for each task to measure the scalability of the visualizations. In general, the more time-series are presented [75, 100], the more difficult the task becomes. Considering the complexity of the visualization designs, we restrict ourselves to four time-series to prevent visual overload. A small pilot study within the working group showed that the simultaneous display of more than four time-series (combined with the large value ranges) resulted in unreadable visualizations due to clutter and lack of available pixels on desktop screens.

For many of the visualizations and tasks, it is easier to compare values when they have a different exponent as differences become more pronounced. Hence, we consider both the case that the exponents $e$ at the target time points for the task are the same in all time-series, and where they have a difference of at least 1.

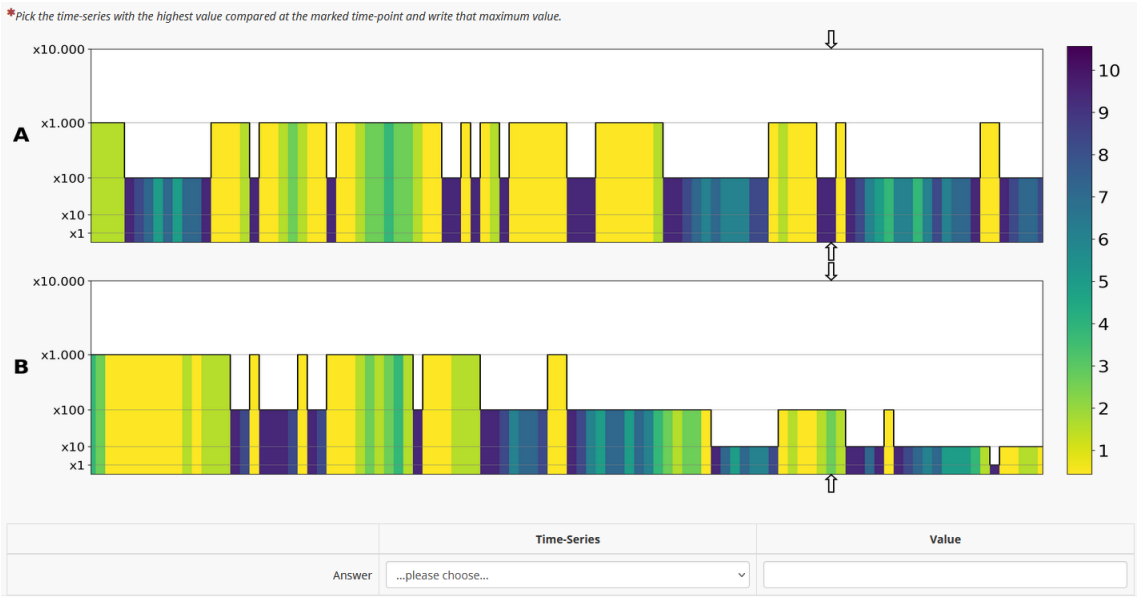



FIGURE 3: Study Interface. The example shows a maximum task for HSLC design. Participants pick the time-series with the maximum value at the marked time point from a drop-down list and type the corresponding numerical value.

**Maximum – *Value Identification and Reading*** Participants had to identify and estimate the time-series containing the maximum value at a specific time point. The target time point is the same for all (2 or 4) series. An example of the interface for the experiment is shown in Figure 3. Here, participants used a drop-down list to select the time-series with the maximal value, and manually typed the value in a text box.

**Discrimination – *Value Comparison*** Participants had to pick the time-series containing the maximum value at *different* time points for each time-series. These time points were distributed randomly through the x-axis. Similar to the previous task, participants selected their answer from a drop-down list. It was not asked for a numerical value.

**Estimation – *Difference Determination*** Participants had to pick the time-series with the minimum value, maximum value, and estimate the value difference between the two time-series. The target time point was the same for all series, reflecting the real-word task of estimating the value range covered by all time-series at a given time. The time-series selection was done via drop-down lists and the value difference had to be typed in a text box.

**Slope – *Trend Detection*** Participants were asked to select the time-series with the highest value decrease over the entire time span. That is, find the time-series with the greatest value difference between the first and last time point. All data sets in the task featured a decrease between start and end to ensure equal difficulty across the trials. As in previous tasks, time-series selection was done via a drop-down list.

**Minimum – *Value Identification and Reading*** The task has the same structure as the maximum task, except participants are asked for the minimum at a specific time.

### 4.2 Data

To generate the data for our tasks, we mirrored the process of Braun et al. [25]. The data for all time-series use generic, contextless variables in order to provide generalizable results. We keep the data homogeneous in terms of scale, i.e. all time-series use the same axis value range. The data covers a value range of $[0, 100\ 000]$ using only integer mantissa. Although value ranges can be arbitrarily large, we have chosen to fix it at this level. Particularly in combination with multiple time-series, displaying larger value ranges even with designs developed for it is difficult due to the limited number of vertical pixels available. In addition, data that exceed this range are often scaled to smaller ranges (e.g., values are shown "in billions" and only the first digits are visualized). Since we do not focus on long time-series or streaming application, each data set consists of 100 data points.

We use constrained random walks to generate synthetic data sets for our user study using a uniform distribution to model ecologically valid data sets [1, 185]. We start with a random value within the range of $[1000, 10\ 000]$. Iteratively, the mantissa is then randomly changed by $[-2, 2]$ to generate a single time-series for a task. The process was then repeated as often as the number of time-series required. Each task and difficulty level used a different set of time-series.

The marked time points were randomly selected from the previously generated data sets according to the conditions of task and difficulty level. We used rejection sampling for the generated data sets and time points to match the conditions (i.e. not decreasing from start to end, or not having time points with the same exponents). The different visualizations designs used the same data per trial for comparability. Although smoothed generated walks may not fully emulate real-world time-series, they enable us to maintain control over study conditions and ensure consistent difficulty levels across all trials. Furthermore, their utilization for data generation has been established in previous time-series research [90, 98, 99].

### 4.3 Stimuli Design

To optimize data readability we followed state-of-the-art visualization guidelines [61, 129, 193]. As we are testing low level tasks such as reading and comparing values in multiple time-series, we use minimalistic presentations. Hence, we exclude any visual embellishments that are not necessary for understanding the visualizations and would distract from the actual data [16, 72]. Similar to Braun et al. [25], we established common design criteria and applied them consistently to all tested designs to ensure consistency in the evaluation.

**Labels:** We excluded labels for the time axis as they are not required for any of the tasks and could cause semantic distractions.

**Grid lines:** Following literature on grid lines [15, 70, 175], we show only a limited number of grid lines. In particular, the main grid lines for the orders of magnitude have tick marks and labels, and a minor grid line with reduced opacity for the mantissa value of 5 is shown.

**Space:** We allocated an equal amount of total screen size (1400x600px) for juxtaposed and superimposed visualization designs. This reflects real-world conditions where the amount of space for a visualization is fixed.

**Juxtaposition:** We positioned all juxtaposed charts from top to bottom to align the time-series. We allow for spacing between the charts to prevent confusion and make room for visual markers.

**Visual markers:** To indicate the time points of interest for relevant tasks, we use visual markers. Previous studies for single time-series used small vertical lines above

and below the visualization [25, 90, 98, 99]. This is not feasible for multiple time-series, as there is no clear indication which time-series they refer to. Therefore, we mark the time point of interest using arrows above and below all charts (see Figure 3) and place these outside the visualization to prevent occlusion.

### 4.4 Experimental Setting

**Procedure** Our two studies were conducted in a **between-subject** setting. Each visualization design defined one of the seven between-subject groups. This study structure provides two major advantages [38]: First, it rules out the possibility of learning effects that occur in within-subject studies. Second, it reduces the total number of trials per participant, allowing us to test for multiple levels of complexity per task.

For each task, the participant was shown either *two or four time-series* on the screen, with the values at time points of interest in these series either having the *same or a different exponent*. This results in a total of 1[design] × 4[tasks] × 2[number of series] × 2[exponent condition] = **16 trials** per participant for the first study and 4 trials for the follow-up study only containing the minimum task. The order of the tasks, amount of time-series shown, and exponent condition were all randomized for each participant to prevent ordering effects. Since we used the same data for all designs, 20 different data sets were generated for the studies.

After each task, participants had the opportunity to take a short break if necessary. The average completion time of the first study was 24 minutes and participants were compensated by £3.18. For the follow-up study, the average completion time was 11 minutes and participants were compensated by £1.75.

Within the study interface, each page hosted a single trial and participants were required to manually click a button to progress to the subsequent page. We omitted a back button to prevent learning effects by subsequent changes of responses. We recommended a screen size of 13" or larger to be able to view the entire trial at a glance (we could not check participants' actual screen sizes in our current system).

At the start of the studies, we asked the participants single-choice questions about their gender, age, degree, and experience with time-series visualizations. The randomly assigned design was then explained to the participants and a practice example including feedback was given on how to read a value in this visualization. Before each task, an additional example and training was provided to explain how to use the respective visualization for the task. In between the tasks, attention checks were included. At the end of the studies, we asked participants about their subjective perceived task difficulty and the aesthetics of the visualization design they saw in the study on a 5-point Likert scale [186]. They were also able to provide free text feedback.

**Participants** The studies were conducted online using Limesurvey [115]. We recruited the participants for our study from the crowdsourcing platform Prolific [145] and reached our target sample size of 105 participants (15 participants per design) in both studies after filtering for exclusion criteria. To participate in the studies, people had to be older than 18 years, speak fluent English, and not have color vision deficiencies, which we tested with the Ishihara tests for color blindness [48]. In addition to the preregistered criteria, we also excluded participants who might not have come into contact with exponential notation before. Hence, we excluded participants without at least a bachelor degree in subjects where one would encounter and use exponential notation.

Out of the 105 participants, 49.5% were female and 50.5% were male in both studies. The age was distributed between 20 and 50, with the majority of the participants (81.7% in the first study, 75.2% in the follow-up) between 20 and 30 years old.

A between-subject study design inherently poses the risk of uneven distribution of expertise among participants. We tested for dependence of the self-reported experience with time-series visualizations using a chi-squared test, which showed no statistical evidence that designs and expertise were correlated (first study: $p = 0.486$, $\chi^2 = 17.54$; follow-up: $p = 0.506$, $\chi^2 = 23.24$). Thus, the study results are deemed comparable.

## 5. Results

For each study task, we measure error (inaccuracy) and response time in total and separated for the number of time-series and exponent conditions tested. In addition, we analyze participants' self-reported task difficulty and perceived aesthetics of the design they had to use.

### 5.1 Analysis

**Error Measurement** In our study tasks, we have two different types of response data: Quantitative responses for the maximum, minimum, and estimation tasks (i.e., the value), and categorical responses in discrimination and slope task for selecting the correct time-series (i.e., the selected series). Therefore, we use two different error definitions to test for inaccuracy.

For the quantitative responses, we use an error measure similar to Braun et al. [29]. While log-errors used in previous works (e.g., as in [94]) either favor over- or underestimation of the sought value and emphasize larger absolute errors, our measure leads to comparable results in terms of values. Moreover, it allows to gain more insights into types of error specific to data with large value ranges: It highlights whether mantissa or exponent errors were made. To calculate the error, let $v_{max}$ be the maximum value and $v_{min}$ the minimum value of the correct and given answer. Let $e(v)$ and $m(v)$ be the exponent $e$ and mantissa $m$ of a value $v$. Then the error is:

$$\textbf{error} = (e(v_{max}) - e(v_{min})) + \frac{(m(v_{max}) - m(v_{min}))}{9} \tag{1}$$

This metric measures the deviation of the response value from the true value in mantissa steps. For example, with a true value of 200, a response of 500 would result in error = 0.33, as would a response of 80, with both being 3 mantissa steps off. Lower values indicate better performance and an error $\geq 1$ indicates an exponent error (in the example above response values of 20 or 2000). This error definition allows for an easy interpretation of the error values:

- $error < 1$: Mantissa incorrect, exponent correct
- $error \geq 1 \in \mathbb{N}$: Mantissa correct, exponent incorrect
- $error \geq 1 \in \mathbb{R} \setminus \mathbb{N}$: Mantissa incorrect, exponent incorrect

For the measurement of errors in the selection of time-series, only binary results of correct or incorrect selections are possible. Hence, we use the percentage of incorrect answers as our error metric for these tasks.

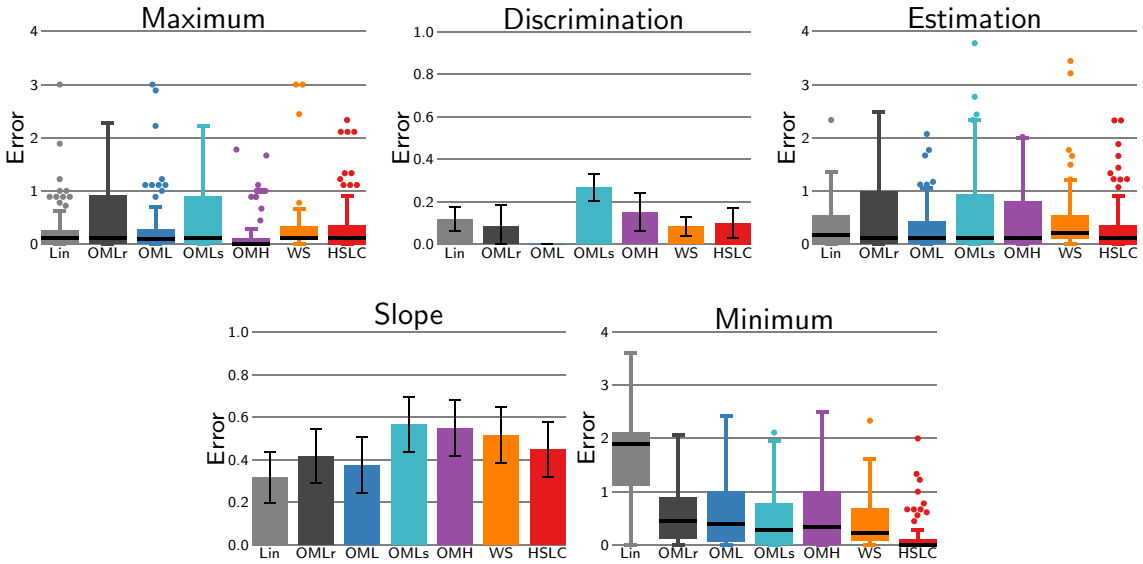



FIGURE 4: Errors per design and study task (the lower, the better).

**Significance Tests** In the analysis of the results, a three-stage significance testing approach was employed for each task, following the preregistered testing plan. Initially, we ran a Shapiro-Wilk test on both error rates and response times that revealed no normal distribution in the data. Thus, the use of non-parametric tests for subsequent analysis is required.

In the second phase of the analysis, different tests were used depending on the type of response data. The categorical responses for the self-reported difficulty and aesthetics were tested for independence using a chi-squared test. The quantitative errors and response times were tested for difference in distribution using the Kruskal-Wallis test. Subsequently, post-hoc analyses were conducted on tasks and responses deemed significant in the preceding stage. Pairwise chi-square tests for categorical responses and paired Wilcoxon signed-rank tests for quantitative responses allow for comparison of all design combinations and testing of our hypotheses. We used two-sample Wilcoxon tests for differences in error and response time for the difficulty levels per design.

All tests were performed with a standard significance level of $\alpha = 0.05$, and a Bonferroni correction factor of 21 was applied to account for the numerous pairwise comparisons per task and measured aspect.

### 5.2 Error Rates

Figure 4 provides an overview of the error rates per task and design, whereas Figure 5 differentiates these errors by the number of time-series and Figure 6 by the different data conditions. All box plots (used for the quantitative answers in the maximum, minimum, and estimation tasks) show the *25% and 75% quantiles* of the data, while the thicker line within each box indicates the median. The whiskers have a maximum length of $1.5 \times$ [box height], but only extend to the furthest data point within this range. In the discrimination and slope task, participants only had to select the correct time-series. For these tasks, the bar charts show the mean error, and the error lines show the error's *95% confidence interval*.

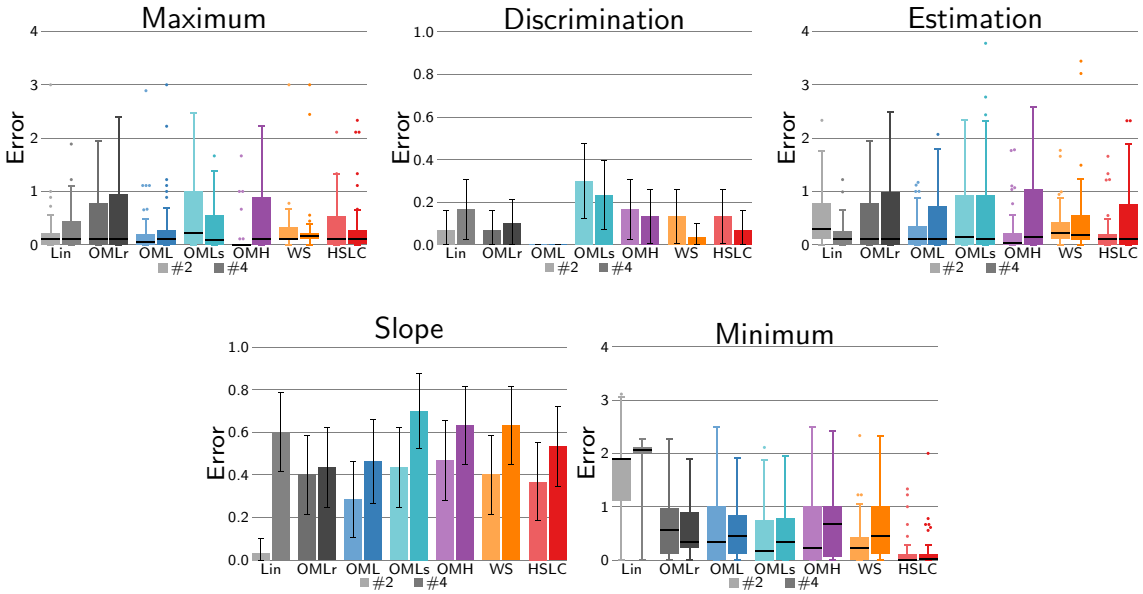



FIGURE 5: Errors per design and study task separated by number of series (the lower, the better).

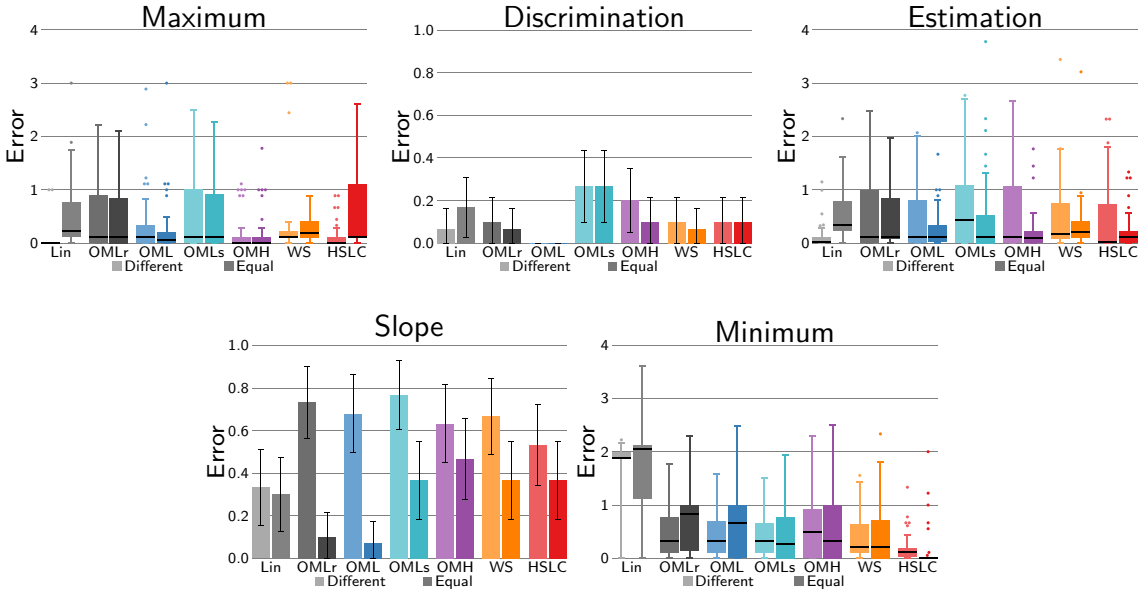



FIGURE 6: Errors per design and study task separated by exponent condition (the lower, the better).

**Maximum Task** For the maximum task, OMH lead to the lowest median error rate ($\tilde{x} = 0$) when reading the maximum value. All other designs shared a median error rate of 0.11. The Kruskal-Wallis test indicated a significant difference in the error rates ($p = 0.01$, $\chi^2 = 15.96$). The pairwise Wilcoxon test showed one significant difference:

- OMH had a lower error rate than WS ($p < 0.01$).

Differentiating by the conditions, two additional significant effects were found by Wilcoxon tests:

- OMH had a higher error rate for four time-series ($p = 0.02$).
- HSLC ($p < 0.01$) and Lin ($p < 0.01$) had lower error rates when exponents differed.

**Discrimination Task** With the OML design, participants made no errors at all in the discrimination task. All other designs except OMLs ($\bar{x} = 0.27$) had an error rate around 0.1. The Kruskal-Wallis test indicated a significant difference in the error rates ($p < 0.01$, $\chi^2 = 22.76$). The post-hoc pairwise comparison showed:

- OML has fewer errors than OMLs ($p < 0.01$).

For the number of time-series and exponent conditions, no significant effects were found.

**Estimation Task** The OMC warming stripes ($\tilde{x} = 0.21$) and the linear line chart ($\tilde{x} = 0.17$) had the highest error rates in the estimation task. Similar to the maximum task, all other designs had a median error rate of 0.11. However, OML, OMH, WS, and Lin have larger variances in the reported answers. For this task, the Kruskal-Wallis test indicated no significant difference in the error rates ($p = 0.19$).

Comparing the error rates per condition, significances were found only for the linear line chart:

- Lin had a lower error rate for four time-series ($p < 0.01$).
- Lin had a lower error rate when exponents differed ($p < 0.01$).

**Slope Task** The slope determination task was the most prone to errors out of all tasks. Lin ($\bar{x} = 0.32$) and OML ($\bar{x} = 0.38$) had the lowest error rates, while error rates exceeded 0.5 for OMLs, OMH, and WS. This means that more than half of the answers given for these three designs were incorrect. The Kruskal-Wallis test showed no significant main effect ($p = 0.05$) between the designs.

The slope task is the only task in which the increase in the number of time-series led to the expected increase in error rate for all designs. The following significant differences were found by Wilcoxon tests for the conditions:

- OML ($p = 0.04$) and Lin ($p < 0.01$) had more errors for four time-series.
- OML ($p < 0.01$), OMLr ($p < 0.01$), OMLs ($p < 0.01$), and WS ($p = 0.02$) had fewer errors when exponents were equal.

**Minimum Task** Our novel design HSLC had the lowest median error rate for reading the minimum value ($\tilde{x} = 0$). OMH and OMLs share the second lowest error ($\tilde{x} = 0.22$), while participants made more mistakes with the order of magnitude line chart designs ($\tilde{x}_{\text{OMLs}} = 0.28$, $\tilde{x}_{\text{OMLr}} = 0.44$, $\tilde{x}_{\text{OML}} = 0.56$). The highest error rate was for the linear line chart ($\tilde{x} = 1.89$). The value 0 was given as the answer for Lin in 82% of the 60 trials. The Kruskal-Wallis test indicated a significant difference in the error rates ($p < 0.01$, $\chi^2 = 143.35$). The pairwise Wilcoxon test showed the following significant differences between designs:

- Lin has a higher error rate than all other designs ($p_{\text{All}} < 0.01$).
- HSLC had a lower error rate than all other designs except OMLs ($p_{\text{OML, OMLr, WS}} < 0.01$, $p_{\text{OMH}} = 0.02$).

Differentiating by the number of time-series and exponent condition, two significances were found by Wilcoxon tests for minimum estimation:

- HSLC had a lower error rate when exponents were equal ($p < 0.01$).
- Lin had a higher error rate for four time-series ($p < 0.01$).

**Summary** Hypothesis **H1** (*HSLC reduces error rates compared to state-of-the-art designs*) is rejected for all tasks except the minimum task, as some designs performed better than HSLC in the others. **H3** (*OMH outperforms all variants of OML in terms of accuracy*) is rejected, as OML had lower error rates than OMH in discrimination, estimation, and slope task. **H5** (*OML charts are more accurate with the use of color*) is partially accepted: OML yields fewer errors in all tasks compared to OMLr, but not significantly. **H7** (*Lin is less accurate than designs for large value ranges*) is accepted only for the minimum task. Lin performed better (non-significant) than all other designs in the slope task and was not significantly worse in the other tasks.

### 5.3 Response Times

An overview of the response times (per trial) per task and design is given in Figure 7. Figure 8 and Figure 9 separate the response times by the number of time-series and exponent conditions. The response times always increase when the number of time-series increases from 2 to 4.

**Maximum Task** Comparing the response times for the maximum task, participants responded the fastest with the linear line chart ($\bar{x} = 22.05$sec), followed by OML ($\bar{x} = 22.9$sec) and OMLs ($\bar{x} = 26.16$sec). The longest time was taken with OMH ($\bar{x} = 34.53$sec) and OMLr ($\bar{x} = 36.87$sec). The Kruskal-Wallis test indicated a significant main effect ($p < 0.01$, $\chi^2 = 45.91$) and the pairwise Wilcoxon test showed the following significances:

- Lin was faster than OMLr ($p < 0.01$), OMH ($p < 0.01$), and WS ($p < 0.01$).
- OML was faster than OMLr ($p < 0.01$) and OMH ($p = 0.02$).
- OMLs was faster than OMLr ($p = 0.01$).

For the conditions, all designs were slower with four time-series. For all designs except HSLC the response times increased when the exponents were equal. Two significances were found by Wilcoxon tests:

- OMLr was slower for four time-series ($p = 0.02$).
- OMLr was slower for equal exponents ($p < 0.05$).

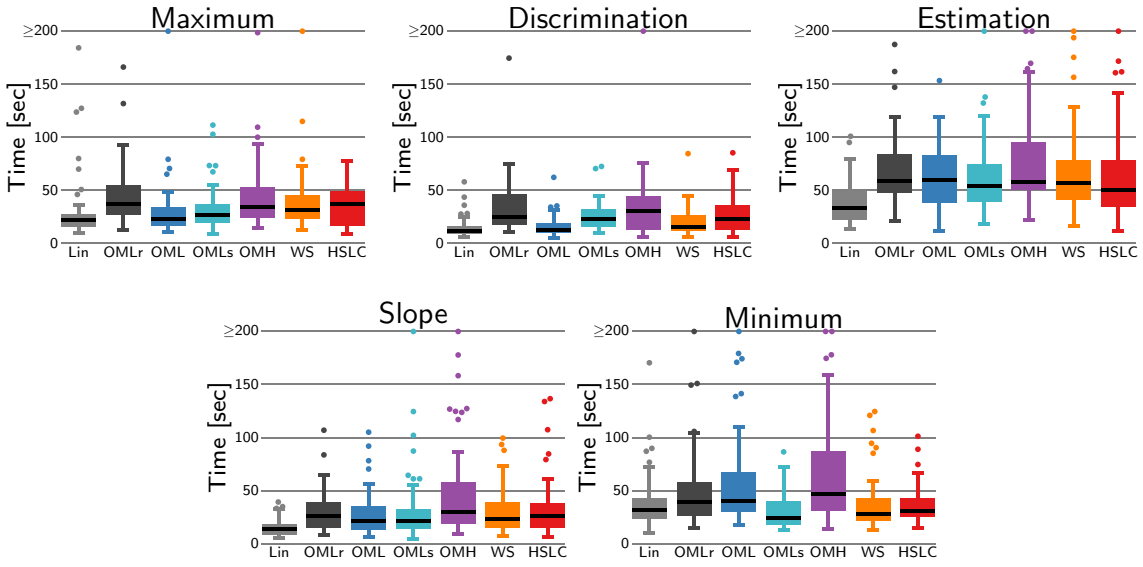



FIGURE 7: Participants' response time per task and design (the lower, the better).

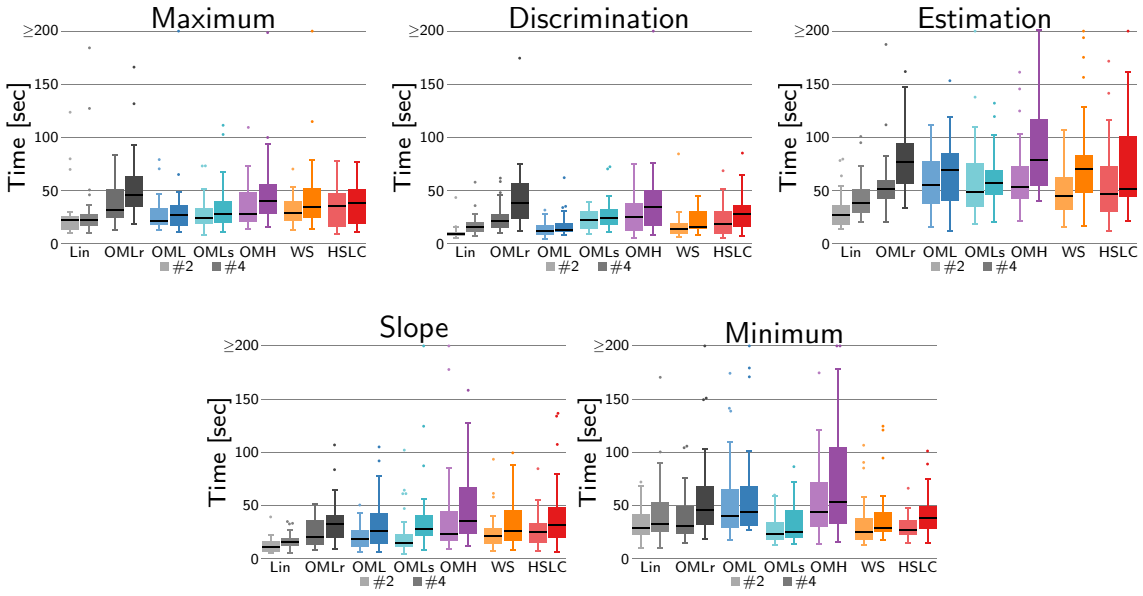



FIGURE 8: Participants' response time per task and design separated by number of series (the lower, the better).

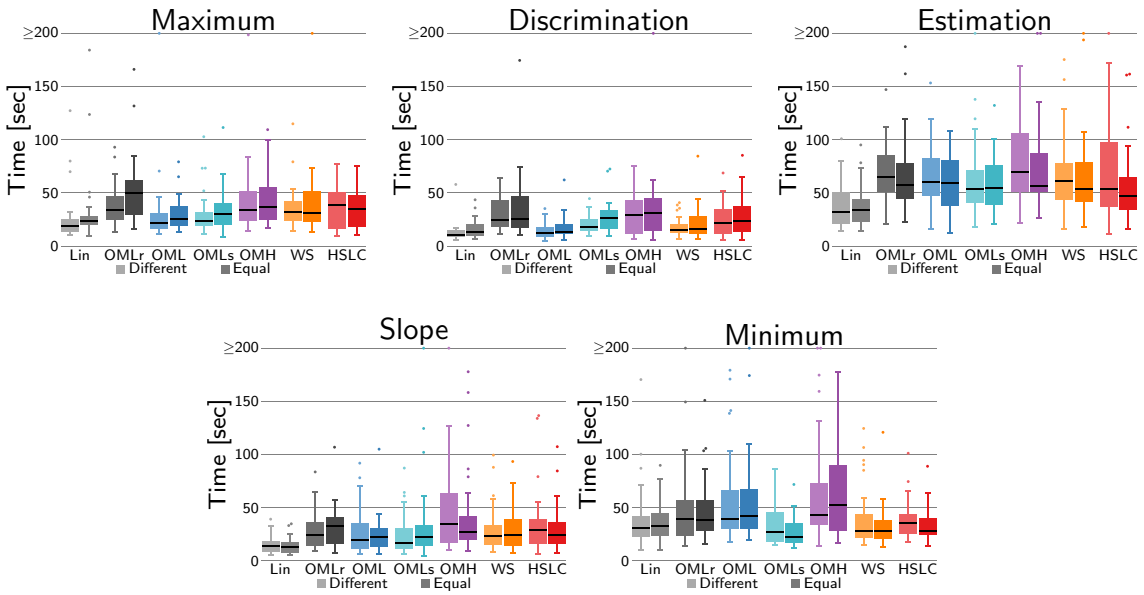



FIGURE 9: Participants' response time per task and design separated by exponent condition (the lower, the better).

**Discrimination Task** For the discrimination task, Lin ($\bar{x}$ = 11.24sec) and OML ($\bar{x}$ = 12.53sec) were again the fastest, and OMH ($\bar{x}$ = 30.43sec) and OMLr ($\bar{x}$ = 24.58sec) the slowest. Kruskal-Wallis indicated a significant main effect ($p < 0.01$, $\chi^2 = 83.41$) and the pairwise Wilcoxon test showed the following significant differences:

- The linear line chart was faster compared to all designs ($p = 0.04$ for WS, $p < 0.01$ for others) except OML.
- OML was faster than all designs ($p_{\text{HSLC}} = 0.01$, $p < 0.01$ for others), except WS and Lin.
- WS was faster than OMLr ($p < 0.01$), OMLs ($p < 0.01$), and OMH ($p = 0.01$).

For the conditions, all time-series were slower for four time-series. For the exponent condition, only OMLs had a substantial jump in the median response times. The following significant results were found by Wilcoxon tests:

- HSLC ($p = 0.04$), OML ($p = 0.04$), OMLr ($p < 0.01$), WS ($p = 0.02$) and Lin ($p < 0.01$) were slower with four time-series.

**Estimation Task** People took the longest time in the estimation task. Again, participants responded fastest with the linear line chart ($\bar{x}$ = 32.98sec). All other designs had similar response times of just above 50 seconds. The Kruskal-Wallis test indicated significant differences ($p < 0.01$, $\chi^2 = 54.69$) and the pairwise Wilcoxon test showed the following significant differences:

- Lin was faster than all other designs ($p < 0.01$ for all).

For the conditions, we again see jumps in the median response time for all designs but HSLC. In contrast to the other tasks, the participants were faster with values in equal exponents for all designs but Lin. The following significant results were found by Wilcoxon tests:

- OMLr ($p < 0.01$), OMH ($p < 0.01$), WS ($p < 0.01$), and Lin ($p < 0.01$) were slower with four time-series.

**Slope Task** The linear line chart also had the fastest response time in the slope task ($\bar{x}$ = 13.74sec). While OMH was the slowest ($\bar{x}$ = 29.5sec), the other designs were in a similar range between 20 and 26 seconds. The Kruskal-Wallis test indicated significant differences ($p < 0.01$, $\chi^2 = 57.32$), and the pairwise Wilcoxon test showed the following significant differences:

- Lin was faster than all other designs ($p < 0.01$ for all).

For the conditions, similar to the other tasks, response again increased for all designs with increased number of time-series. For equal exponents, HSLC, OML, OMH, and Lin were faster while the opposite was true for OMLr, OMLs, and WS. The following significant result were found by Wilcoxon tests:

- OML ($p < 0.01$), OMH ($p = 0.04$), Lin ($p = 0.02$) were slower with four time-series.

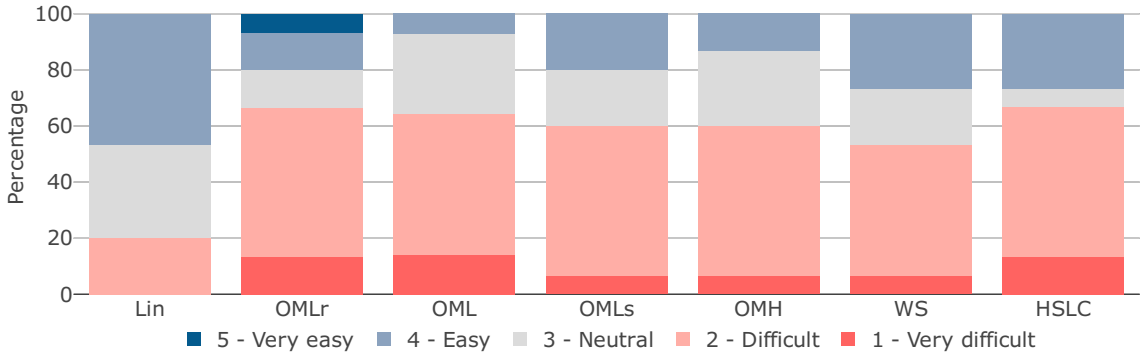
**Minimum Task** For the minimum task, OMLs got the fastest response times ($\bar{x}$ = 30.55sec), followed by HSLC ($\bar{x}$ = 35.94sec), WS ($\bar{x}$ = 36.71sec), and Lin ($\bar{x}$ = 38.68sec). OMH ($\bar{x}$ = 69.78sec) and OML ($\bar{x}$ = 58.39sec) were slowest. The Kruskal-Wallis test indicated a significant main effect ($p < 0.01$, $\chi^2 = 46.09$) and the pairwise Wilcoxon test showed the following significances:

- OMLs was faster than OMH ($p < 0.01$), OML ($p < 0.01$), and OMLr ($p < 0.01$).
- HSLC was faster than OMH ($p < 0.01$) and OML ($p = 0.01$).
- WS was faster than OMH ($p < 0.01$) and OML ($p < 0.01$).
- Lin was faster than OMH ($p = 0.02$).

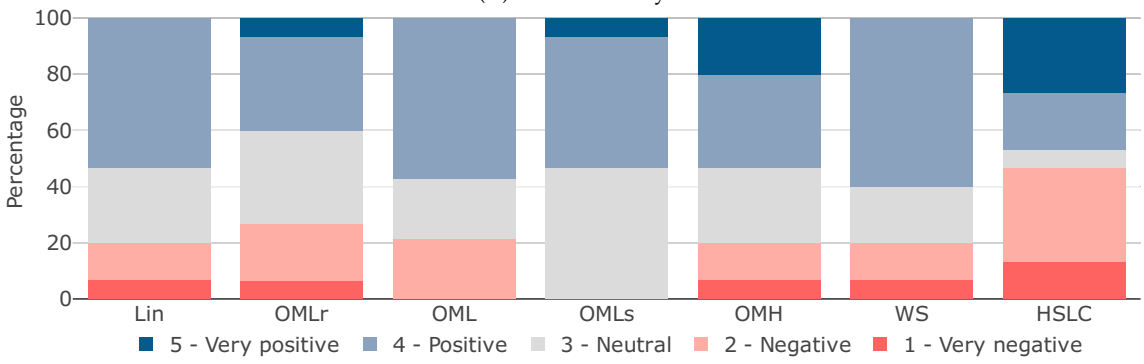
For the conditions, all designs were slower with four time-series, while the result did not differ much for the exponent condition. Wilcoxon tests found two significances:

- HSLC ($p < 0.01$), OMLr ($p = 0.03$) were slower for four time-series.

**Summary** We reject hypothesis **H2** (*HSLC reduces response times compared to state-of-the-art designs*) as HSLC design was not among the fastest in any task of the first study. We also reject **H4** (*OMH outperforms all variants of OML in terms of response time*) since participants had faster response times with OML than with OMH in all tasks. **H6** (*OML charts are faster with the use of color*) is partially accepted as OML with color had lower response times than OMLr in all tasks (significant for maximum and discrimination) except the minimum task. **H8** (*Lin is slower than designs for large value ranges*) is rejected as Lin had the fastest response times for all tasks except minimum identification (significant for almost all).




(A) Task difficulty.




(B) Aesthetics.

FIGURE 10: Distribution of participants' self-reported task difficulty and perceived aesthetics per design for the first study.

### 5.4 Difficulty and Aesthetics

We asked the participants to rate the aesthetics of the designs as well as their perceived task difficulty on a 5-point Likert scale [186]. Figure 10 presents the given responses for the first study. We perform chi-squared tests to determine whether there were aesthetic preferences for any particular design or correlations between designs and task difficulty.

The tasks were perceived as easiest with Lin ($\bar{x} = 3.27$), while the average reported difficulty for all other designs was about 2.5. Participants rated the superimposed OML chart as the most aesthetically pleasing ($\bar{x} = 3.6$), while the average responses for aesthetics for all other designs were between 3 and 3.5. Nevertheless – for both criteria – no significance could be found ($p_{\text{difficulty}} = 0.58$, $p_{\text{aesthetics}} = 0.17$).

In the second study, there were no differences in responses for design aesthetics. For difficulty, there was only a difference for the minimum task. Here, HSLC and WS were perceived easier, and Lin was perceived more difficult for the minimum task than for other tasks.

### 5.5 Free Text Feedback

Overall, 66 participants gave free text feedback on our studies. 15 participants stated that they liked the study and that our explanations of the designs and tasks were helpful. They had the most difficulties with the minimum and estimation tasks, and reported that the tasks increased in difficulty with four time-series due to the reduced chart sizes ("the more graphs the worse it became to figure out the values"). Feedback on individual designs mainly related to color coding. For HSLC, four people suggested a discrete color scheme for a better identification of the mantissa value. We discussed this already during

the design process and decided against it because only integer mantissa could be encoded with a discrete color scheme. With OMH, three people had problems identifying the exponent and suggested higher color differences for the orders of magnitude. For the warming stripes, two participant suggested not to use color alone and to try different visual encodings, e.g., "different patterns". With the line charts without color (Lin, OMLr, and OMLs) "it was hard to read the value and its line" and "it was [...] harder to estimate the coordinate", suggesting that the encoding of values by color supported participants' perception.

### 6. Discussion

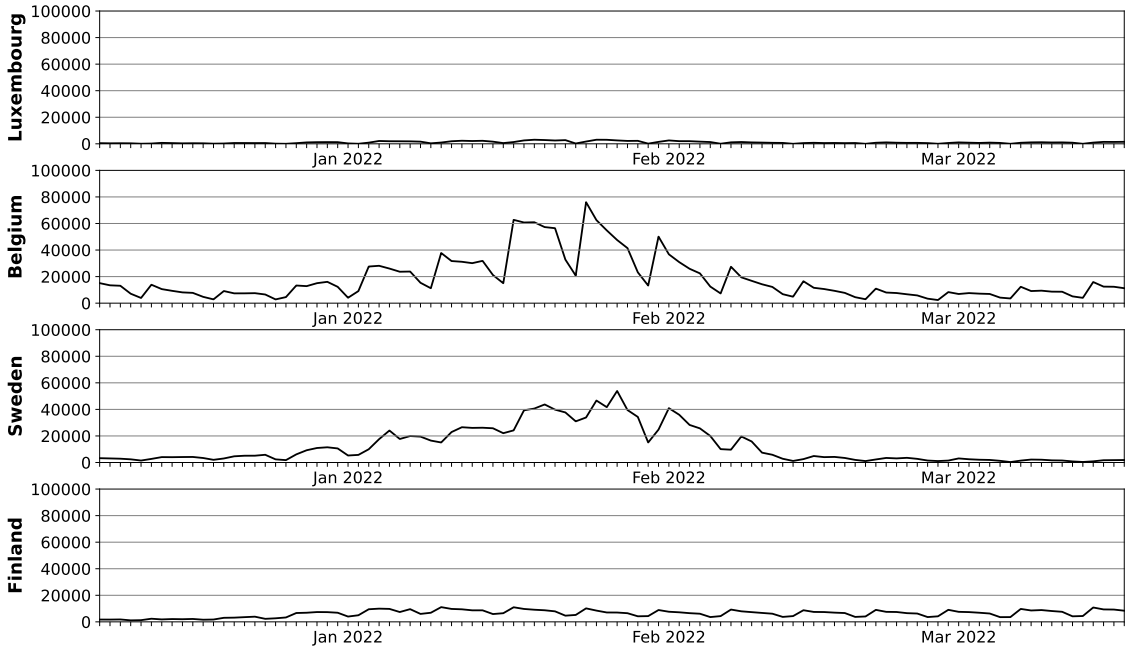
This study evaluated how increased complexity affects readability in two ways: Examining large value ranges instead of small value ranges, and multiple instead of single time-series. Previous studies [23, 25, 29, 93, 94] have shown that visualizations designed for large value ranges improve people's ability to read and compare values in such data. The order of magnitude horizon graph [25] outperformed state-of-the-art as well as standard designs for single time-series. In our study on multiple time-series, no visualization technique consistently outperformed other designs in terms of accuracy. Even the linear line chart, which does not specifically address the characteristics of large value ranges, was not significantly worse than the other approaches in terms of accuracy. With respect to task completion time, it significantly outperformed the other designs, most likely due to familiarity with this design. This is a surprising answer to our research question **RQ1** contrary to previous results on individual time-series. Only for the identification of minimum values the linear line chart was significantly worse due to the small number of pixels available in low orders of magnitude. Here, we would recommend using the HSLC design as it performs best on this task. Figure 11 shows our HSLC design applied to real-world data and compared to the linear line chart.

One explanation for why the linear line chart performed so well, is that it is relatively good for tasks involving the maximum. After all, only small values disappear in a linear plot. For the maximum and comparison tasks, when the exponents differ, the maximum can be used as a perceptual proxy (a large value minus a small value is still a large value). This explanation is supported by comparing the results for different exponent conditions. When the exponents are equal, the maximal is more difficult to read. Thus, errors were significantly worse because the perceptual proxy could not be used (including the minimum identification).

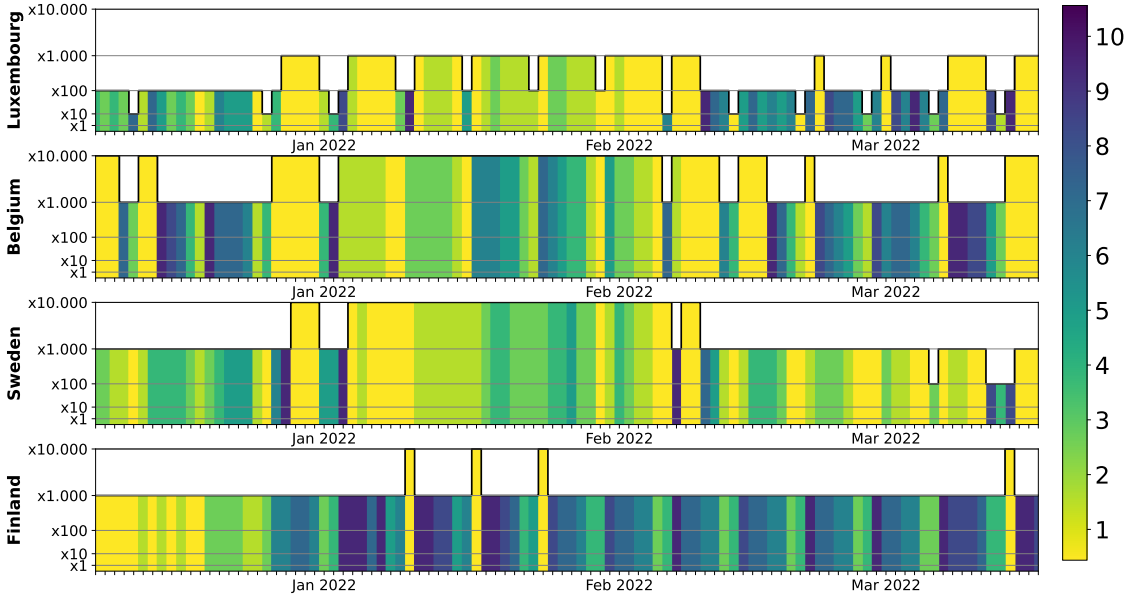
Answering research question **RQ3** and comparing our novel height-stack line chart design to the state-of-the-art designs, its results were average for the tasks in the first study not performing significantly better or worse. For the minimum task, it however outperformed all other designs in terms of accuracy. Therefore, it is an appropriate solution for identifications and comparisons of values across the entire value range including several orders of magnitude.

This decrease in performance of designs that worked well for large value ranges in single time-series, when applied to analysis of multiple time-series, indicates that novelty plus increased complexity of the overall layout may have overwhelmed the participants. Nevertheless, experts may be supported by large value ranges visualizations given that they have more time and incentive to learn how the visualization should be read.

The increase in data complexity (i.e., the addition of large value ranges) provided new insights on multiple time-series. Previous research on multiple time-series with small value ranges [100, 118, 135] showed that chart-wise juxtaposition almost never outperforms other layouts, especially in comparison tasks. In our study, the only superimposed design (OMLs) had the highest error rates in reading tasks and tasks, in which time-series




(A) Linear line chart.




(B) Height-stack line chart.

FIGURE 11: Application example: The number of daily new COVID-19 cases compared for different countries visualized with the linear line chart and our new height-stack line chart.

had to be compared at different points in time. This suggests overlap problems to be amplified with large value ranges. In comparisons of time-series at a fixed point in time, the layout worked well again.

Moreover, we found that the intuitive assumption that the error rate should increase with an increased number of time-series is not true for large value ranges. For maximum, discrimination, and estimation, the error rates decreased with four time-series for several designs. One possible reason for this phenomenon could be that the participants were no longer able to perceive answers at first glance with an increased number of time-series. Therefore, they had to look at the time-series more closely, which led to more accurate responses. Thus, the tasks shifted from a perceptual to cognitive problem. The increased response times in all tasks for four time-series support this suggestion. In general, both increasing the value range and comparing a higher number of time-series results in a significant increase in visual clutter (e.g., by gridlines and tick marks; see Figure 1). Therefore, if space is limited, it is recommended to use designs with less clutter, such as WS to OMH.

The differentiated view of the results based on the difference in exponents revealed different outcomes for the designs. This indicates, that they have different strengths and weaknesses dependent on data characteristics. For the designs intended for large value ranges, the results are equal or even improve when the exponents are the same, except for HSLC on the maximum task. This indicates that these designs work best when the exponents are the same, which is particularly relevant for small value differences, which are otherwise invisible.

The slope task was the only task where participants made consistently more errors when the number of time-series increased (i.e. four vs. two time-series). For all designs except Lin and OML, more than 40% of the given answers were incorrect, which is akin to the the expected error of random guessing. This indicates, that trend perception is very difficult with the adapted axis scalings in large value ranges visualizations, aligning with previous results [25]. For slope tasks, line charts have a significant drop in performance when the complexity of the data increases.

Our study also provides indications on the role of visual mapping of the exponent to color. The order of magnitude line chart lead to faster response time with color than without color for all tasks (except the minimum task, for which the response times were almost equal), and significantly so for three out of five tasks. Additionally, including color lead to fewer errors (although not significantly) on average. This result is further strengthened by the warming stripe design, which maps the exponent to only color. WS did not perform significantly worse than the order of magnitude line chart both with and without color. Thus, the answer to **RQ2** is that color supports the perception of magnitude variations, which clarifies an open question from Braun et al. [25] and confirms the results of previous research [2,51]. A possible reason for this fact is that changing hues of the different exponents, as opposed to a difference in the y-coordinate, made them much easier to perceive with increasing complexity through multiple time-series. Given the constrained screen size per time-series, the color facilitates the perception of magnitude variations to a greater extent than positional changes in juxtaposed settings. However, using color always carries the risk of not being safe for color vision deficiencies. This is particularly problematic with the OMC color scale [29], in which the color number increases with the exponent. Our HSLC design does not have this problem because the color usage is limited to the constant mantissa range.

## 7. Limitations and Future Work

We discuss the results and limitations of our study in terms of the tested designs, study setup and insights. The unexpected results open up new research questions and opportunities for future research.

**Results: Increase of Complexity** By studying large value ranges for *multiple* time-series, there was an increase in complexity compared to previous studies on large value ranges in *single* time-series. The results showed significant differences with respect to the single time-series setting.

Moreover, we counter-intuitively observed that the error rates decreased in several cases as the number of time-series increased. We discussed possible reasons for this phenomenon in the previous section. It would be interesting to further investigate whether the same observations can also be made in other settings. Examples could be multiple, non-temporal, data sets with large value ranges, or large value ranges in graphs (e.g., as nodes or edge weights).

**Results: Designs** The results for the individual tasks revealed that none of the existing state-of-the-art designs outperformed the linear line chart in multiple time-series with large value ranges for tasks, in which the maximum value is sufficient as an approximation. As a result, further research into visualization designs for large value ranges is needed. The slope assessment task exhibited the lowest accuracy across all designs, with participant results nearly akin to random guessing. Therefore, the trade-off between simplified trend detection and adequate readability and comparability of values of all orders of magnitude remains an open problem in the design process for large value ranges.

**Experimental Setup** We conducted *online experiments* as is common in comparable studies. The online setup, however, despite the training phase, does not allow us to ensure that the participants had a full understanding of the visualization techniques. The results showed that some participants were overwhelmed by the new approaches without additional guidance. It would therefore be interesting to replicate the studies in a setting with personal supervision of the participants. This would help to gain deeper understanding behind the task completion processes and their difficulties.

**Scalability** Our studies included trials with 2 and 4 time-series and a value range of $[0, 100\ 000]$. Both have been used similarly in previous works [25, 75, 90, 100] and are ecologically valid. Further increases in both of these factors would require additional research on visual clutter in large value range visualizations.

**Evaluated Designs** While our study has tested seven different designs, improvements for individual designs can still be investigated. We also examined the influence of color on task performance in general, but did not study the particular color choice. Participants suggested that the color selection of the order of magnitude color scheme [29] is not intuitive and that the coloring of the HSLC design could be improved for a better discrimination of the mantissa values. Moreover, the perception of the order of magnitude in OMH could be enhanced by a higher contrast in the respective hues [180]. Additionally, our study was limited to static charts; incorporating interaction may yield better results, especially for more complex visualization techniques [112, 141].

We used homogeneous data in terms of scale and units, i.e., the same axis value range was presented in each individual time-series. Different design considerations would be necessary to compare heterogeneous data. For instance, it would be worthwhile to research the application of indexing techniques [3] to large value ranges visualizations.

## 8. Conclusion

We conducted two empirical studies to compare visualization designs for large value ranges in multiple time-series. We evaluated five state-of-the-art approaches including adaptions for large value ranges, one standard technique for time-series data, and our newly introduced height-stack line chart. Participants had to solve five common tasks in large value ranges and time-series research: maximum and minimum identification, value discrimination, difference estimation, and slope assessment. Our findings suggest that the complexity introduced by combining multiple time-series and large value ranges challenges the findings of previous studies in these areas. For instance, error rates did not increase in several designs as the number of time-series increased, contrary to expectations. Designs proven effective for large value ranges in single time-series performed poorly with multiple time-series. Surprisingly, none of the designs specifically tailored for representing large value ranges significantly outperformed the traditional linear line chart in terms of accuracy and response time for all but the minimum task. Our findings highlight the importance of future research on the cognitive processes involved in solving complex tasks and the design of appropriate visualizations for large value ranges.

## Acknowledgments

The authors would like to thank all study participants and the reviewers, whose suggestions helped improving this paper. This work has been partially supported by BMBF WarmWorld Project and KPA Intelligent Methods for Earth System Sciences.

# Chapter 3

# Visual Validation of Regression Models

## 3.1 Theoretical Framework of Visual Model Validation and Estimation in Visual Analytics Processes

Model estimation (i.e., selecting the model type and its parameters for given data) and model validation (i.e., assessing the quality of a given model) are core components of model building and exploratory data analysis [49, 96].

In visual analytics systems, model building (i.e., feature and model selection, parameter setting, and model validation) [117, 158] can take place in workflows where model estimation and validation are performed in a sequence of multiple steps. They can be performed in the same computational or visual way or can vary between the steps. For example, a computer-generated model can be first computationally validated and then visually validated [42, 130]. Or a computationally estimated model can trigger a visual validation and subsequent visual model adjustment. A common example is time-series modeling: An analyst needs to see the time-series graph before choosing a suitable modeling method. After creating an initial model, the analyst validates it by comparing it visually with the data. Model estimation could be done interactively in an intelligent system, where the analyst does not explicitly select the method and set its parameters, but instead sketches what the model prediction should look like. The system then finds an appropriate model and fits it to the data. The explicit support of such visual and interactive model estimation and validation is more user-friendly, decreases the cognitive load, and can be used by non-experts. Thus, they are essential parts of modern data analysis workflows [49, 96] and need to be captured in VA pipelines.

Existing visual analytics frameworks cover a broad range of modeling workflows in VA systems [11, 12, 159, 173, 174, 192]. However, they are often insufficiently described or lack detail about the exact processes and components. Moreover, they do not include which parts of the processes are actually carried out by humans and which by computers. In this section, "model building loops" are decomposed and the roles of visual estimation and validation within these loops are explicitly characterized. Further, the influences exerted by human analysts and automated algorithms are differentiated, and all feasible human–computer interaction patterns that emerge when these elements are combined are enumerated. Building upon the VA pipelines by Keim et al. [107] and Andrienko et al. [12], a novel VA pipeline is proposed that subsumes prior visual-interactive modeling workflows and extends them to integrate both visual model estimation and validation (see Subsection 3.1.2).

### 3.1.1 Visual Estimation versus Visual Validation

To perform **model estimation**, usually the computer derives the best fit to the data at hand given the model parameters. These parameter can be set programmatically or interactively in the interface. The modeling steps or results can be interactively visualized (e.g., the iterative steps of clustering algorithms in the EduClust platform [77]). Visual estimation is performed when a data scientists visually inspects the data first to decide which type of model to calculate or when a chart only shows the data but a corresponding text (e.g., on a website or in newspaper) mentions a relationship. Alternatively, visual model estimation can be performed, when a human looks at the data and directly manipulates the model result on the screen until the model fits the data [46, 52, 80, 95].

Performing **model validation** means to check whether a given model result is a good fit to the data. This model verification can be done computationally or visually. In the first case, the computer calculates model quality statistics and checks whether the values are suitable. In the second case, the model results and data are visualized and the human assesses the fit of data to the model. Visual inspection of complex models to check their correctness and reliability is crucial [40, 41], as statistical metrics are often insufficient to describe the model and data [122, 181].

The main difference between the two processes is the influence of computers and humans on model estimation. While the model parameters are calculated by the computer during computer-based model estimation, the transformation from data to model parameters takes place solely in the user's mind during visual model estimation. This means that with visual estimation, the user takes one step away from the computer in the visual analytics process.

Furthermore, visual estimation contains a validation component, since each time the model is adjusted, a decision has to be made as to whether additional adjustments are needed. In the visual validation task, on the other hand, people compare the shown model to their own mental model. Thus, both processes are not necessarily in opposition to each other (i.e., either one or the other). They can also co-exist, such that the human mental model is used to control the creation and iterative refinement of the computer model, which, in turn, serves as a tool to iteratively improve the human mental model. Therefore, visual validation and estimation are neither purely perceptual nor cognitive and the interplay between these two mental processes needs to be researched in more detail.

In summary, visual model validation and estimation must be considered separately in visual analytics pipelines. However, they are often not included or only implicitly incorporated in such frameworks.

### 3.1.2 Extended Visual Analytics Pipeline

The proposed, novel VA pipeline in Figure 3.1 enhances the frameworks of Keim et al. [107] and Andrienko et al. [12]. In their schemes, visual model validation and estimation are implicitly included in the perception, interaction, and refinement loop of model and visualization, but their differentiation is not clear. The new framework expands their schemes and divides the model component into computer-generated and human-created models. Moreover, a validation component is added, which positions itself between the human perception and knowledge generation. The validation component closes the loop

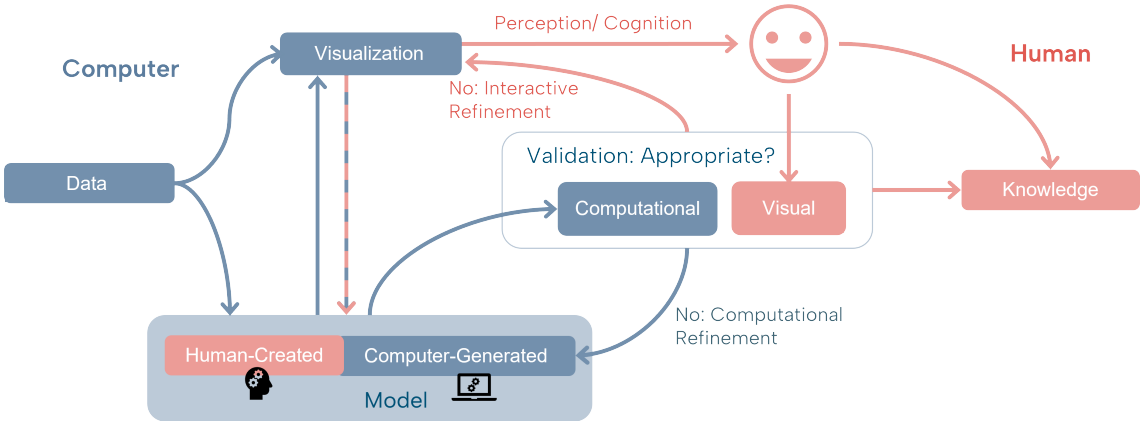


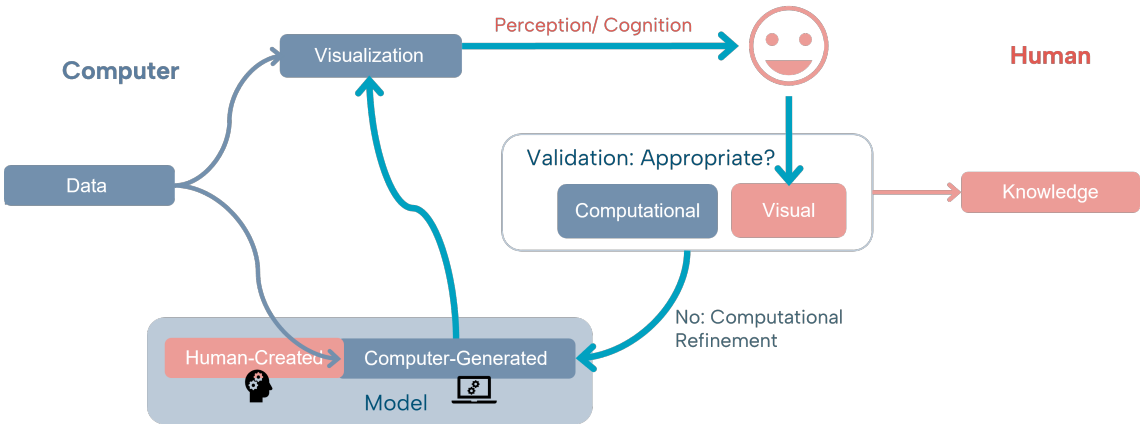

FIGURE 3.1: The proposed visual analytics pipeline with integrated visual model estimation and validation loops (extending [12, 107]).

on model building. From the final model, the user can gain insights into the data. Within the model building cycle, each data model needs to be validated, either computational or by the user. In both visual model validation and estimation, the model is validated visually. Their difference lies in the origin of the models and their re-adjustments.

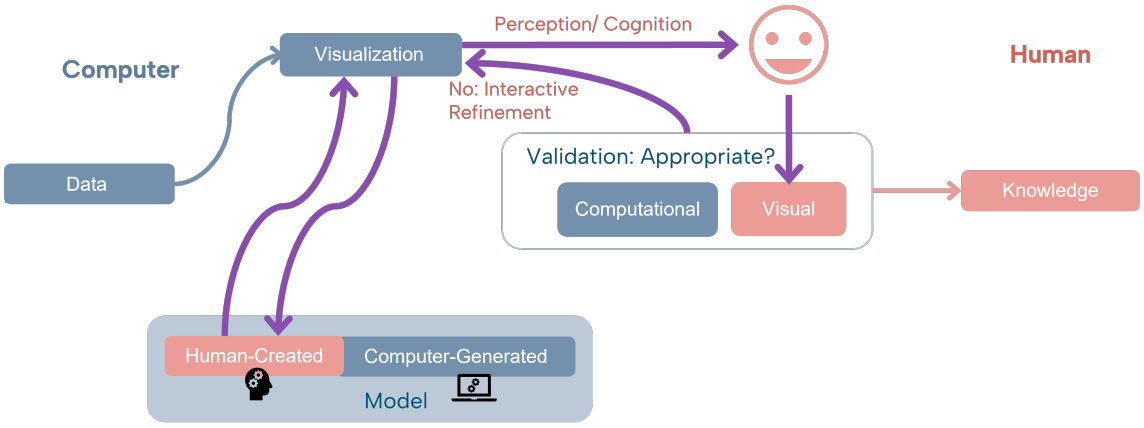
Figure 3.2 shows the explicit visual model validation and estimation loops within the proposed pipeline. In the visual model validation loop, the system automatically generates a candidate model, renders it alongside the raw data, and prompts the user to either accept or reject it. It gives the user an immediate sanity check on the plausibility of the computer's inference and guides the user toward more informative features or transformations by surfacing systematic mismatches between model predictions and observed patterns. In case of rejection, the model parameters are re-generated by the computer, enabling the discovery of alternative hypotheses without requiring the user to tinker with low-level details.

In contrast, the visual model estimation loop gives the analyst direct control over parameters: they iteratively adjust model components interactively and immediately see the impact on the visualization. This manual refinement path is especially useful when the user has a-priori expectations (e.g., known symmetries or constraints) that are difficult to encode algorithmically. Each manual model adjustment is followed by the same visual validation check. By toggling between estimation and validation - accepting a visually satisfactory model or refining it further - the user incrementally creates knowledge about the data and model. The refinement of the model is optional and the process can be stopped at any time.

Together, these loops foster a sensemaking cycle: the validation loop fuels discovery by highlighting discrepancies, while the estimation loop embeds expert knowledge through targeted refinement. Because both loops are optional and freely interleavable — analysts can switch between letting the computer propose models and refining them themselves - the pipeline supports a spectrum of workflows, from hands-off exploration to expert-driven modeling, with the ability to stop once a sufficient model is created.




(A) Visual model validation loop.




(B) Visual model estimation loop.

FIGURE 3.2: The respective analysis processes in visual model validation and estimation tasks in the proposed VA pipeline. Both processes as well as the distribution between human and computer can be changed an intermixed, resulting in an interplay of both tasks.

By explicitly capturing all possible visual and computational loops, the pipeline supports seamless "handoffs" between human and machine - the possibility to switch from human to computer and back - reflecting the iterative, mixed-initiative nature of modern VA systems [22, 36, 42, 96, 130].

For clarity, additional details from previous frameworks [12, 159, 173, 174] have been omitted. These are still relevant and can be added if required, just as the perception part of the process is influenced by bias, cognition, and trust [39, 41]. Thus, the addition of visual model validation and estimation does not limit capabilities of previous visual analytic processes in the new pipeline.

### 3.1.3 Usage and Comprehensiveness of the New Pipeline

The usage and completeness of the new pipeline is demonstrated using example VA systems for regression models [165], dimensionality reduction [101], and decision trees [188]. These VA system include visual estimation and validation and offer many options and interactive loops for creating model results. The examples can be extrapolated to other systems and other types of models, such as clustering, classification, and others [42, 44, 62, 77, 113, 130].

**Regression modeling:** In the interactive regression lens (IRL) [165], firstly, the user looks at the raw data without any model. She interactively selects an area of interest, on which to perform the regression analysis (the so called lenses in Figure 3.3**a**). In order to determine a suitable area, the user must have already made a mental estimate of the model. This corresponds to the human-created model in the new pipeline. In the next step, the visualized regression model can either be automatically generated by the computer, or the user can interactively select the model type and define certain parameters (Figure 3.3**c**). In both cases, the model needs to be validated. This can be done either computationally or visually. For computational validation, various computed regression statistics are displayed, which can be used by the user for validation purposes (Figure 3.3**b**). Visual validation is done by looking at the computed model result line together with the original data set (Figure 3.3**a**). This corresponds to the visual validation loop in the novel pipeline. Subsequently, the model can be refined interactively, e.g., by excluding individual points, switching model types or moving the lenses (see the movement of the activated lens indicated by the arrow in Figure 3.3**a**). This corresponds to the visual estimation loop in the pipeline. As mentioned, visual estimation inherently incorporates visual validation.

**Dimensionality reduction:** At first, the iPCA tool [101] shows the computer-generated model results in Figure 3.4**a**. The included parallel coordinate plots for data and eigenvectors help the users to visually validate these results. A computational validation is done via the correlation view (Figure 3.4**d**). The model can be interactively refined by the user by either manipulating the dimension (Figure 3.4**b**) or specific data items (Figure 3.4**c**). Although the interactions for model refinement are only indirect (i.e., using sliders and buttons), the user acts in the visual estimation loop of the pipeline shown in Figure 3.2b. The user's visual estimation capabilities would be even more powerful if they could interact directly with the model visualizations [32, 67]. Thus, the visual model validation loop is more strongly represented in the iPCA tool.

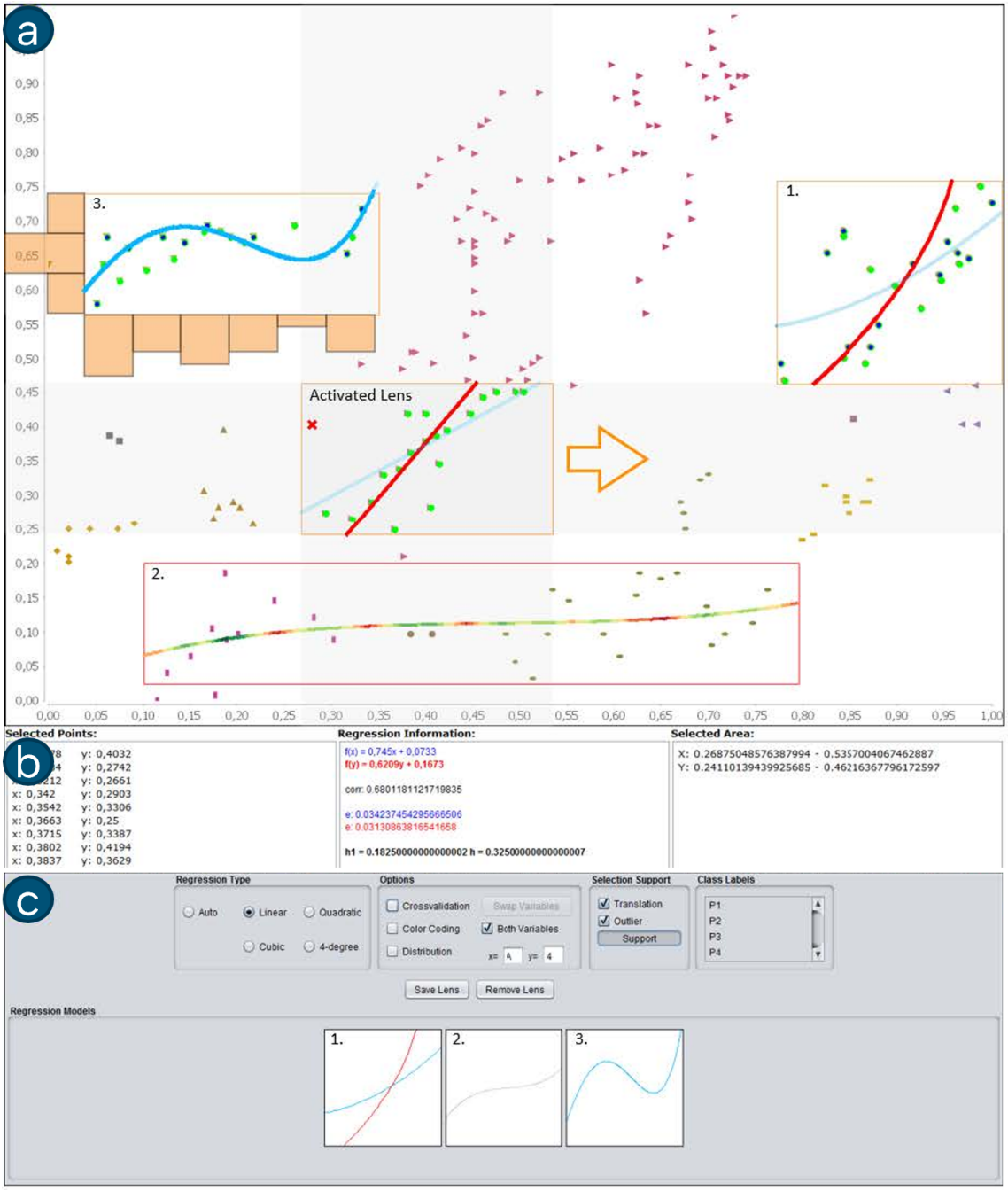



FIGURE 3.3: Shao et al.'s interactive regression lens tool [165].

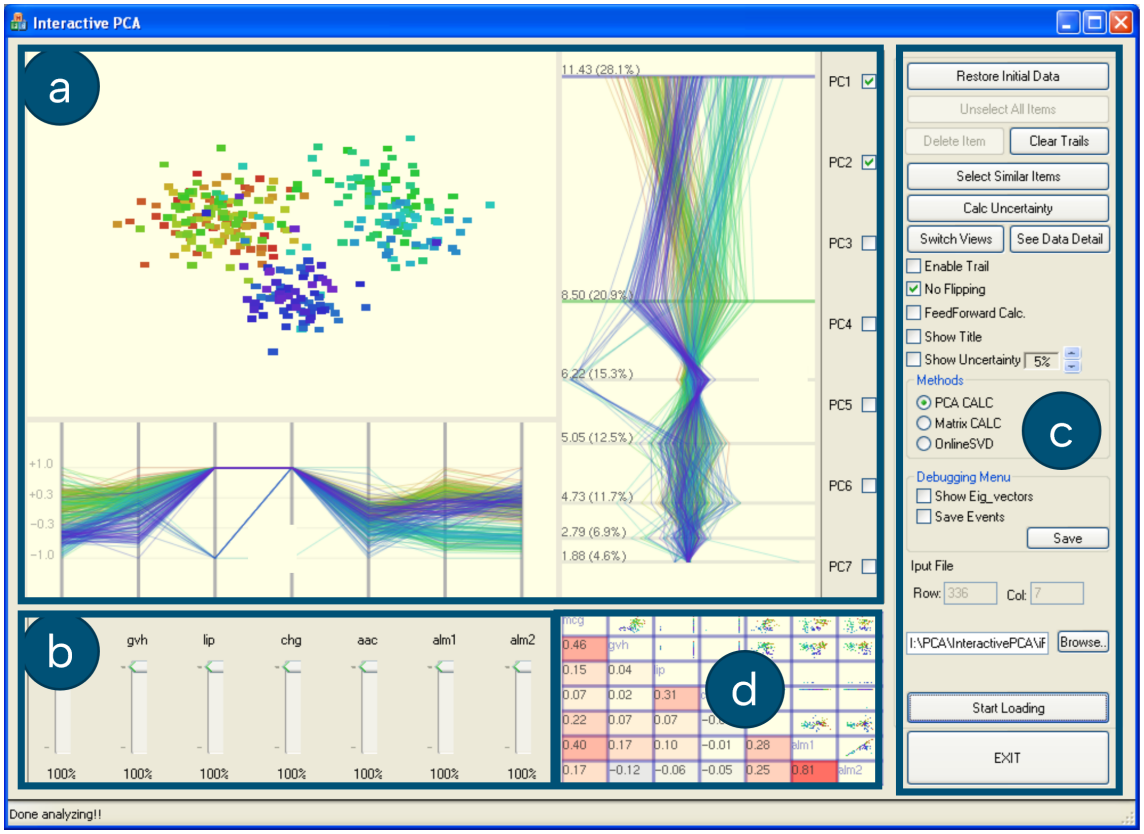



FIGURE 3.4: Jeong et al.'s interactive PCA tool [101].

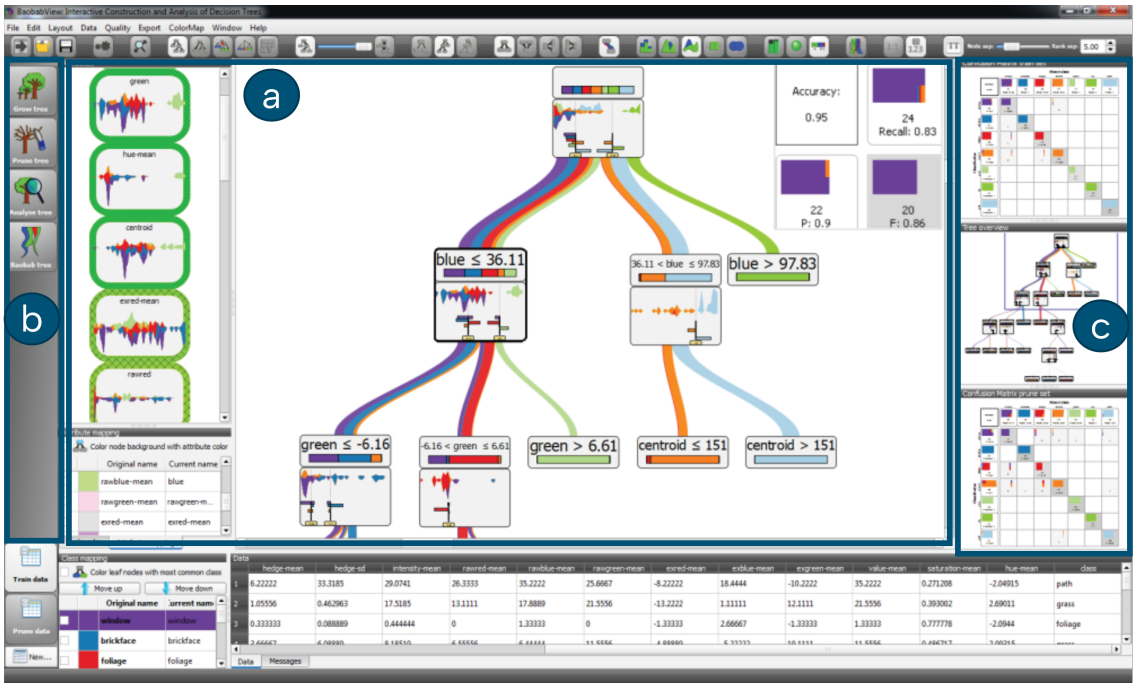



FIGURE 3.5: Van den Elzen and van Wijk's interactive decision tree construction tool [188].

**Decision trees:** A similar workflow structure is visible in the BaobabView [188]. In this system, both of the analysis loops (Figure 3.2) are included and can be performed by the user. The user can manually create a decision tree by interactively defining split points, choosing split attributes or merging nodes. The resulting tree then needs to be visually validated using the tree visualization including additional node information (Figure 3.5**a**). The validation process is algorithmically supported by the confusion matrices in Figure 3.5**c** that suggest misclassifications on the training set. Afterwards, the tree can be further adjusted. This workflow represents the visual estimation loop. Besides the user-based generation of the tree, the system also provides the option to select layouts that are preset by the computer. Followed by a validation of the results, this would correspond to the pipeline's visual validation loop. This loop could be enhanced in the system by providing more options for computational model generation besides providing preset layouts. Therefore, the BaobabView focuses more on the visual estimation rather than the visual validation loop.

*Summary:* The presented systems allow the user to seamlessly switch between the computational and visual estimation as well as computational and visual validation loops. This is all covered in the new pipeline (see Figure 3.2). It demonstrates both the need to include both processes in VA frameworks and the ability of the novel pipeline to cover various types of visual analytics systems to support different modeling processes.

### 3.1.4 Implications and Further Work

Incorporating visual model validation and estimation into visual analytics systems reshapes both their design and interactive workflows. The examples in Subsection 3.1.3 show that designers have to decide which of the discussed modeling loops their systems should support. In best case, novel VA systems include all possible estimation and validation loops.

When a system supports model estimation, it must provide intuitive interfaces for parameter exploration and real-time feedback on model performance. Simultaneously, visual model validation demands clear representations of uncertainty, error distributions, and quality indicators to ensure users can assess model reliability. Moreover, designers must carefully manage cognitive load by integrating validation and estimation views without overwhelming users. Together, these considerations compel visual analytics systems to evolve from passive data explorers to active, model-aware environments that foster human insights and incorporate them into algorithmic solutions.

Understanding the perceptual and cognitive processes involved in visual model validation and estimation is essential for the effective design of visual analytics systems. VA systems are intended to support human reasoning and decision-making by leveraging visual representations of complex data and models. However, their effectiveness critically depends on how users perceive, interpret, and mentally process these visualizations.

In the remainder of this chapter, the perceptual and cognitive differences associated with the two tasks are studied. This analysis is grounded in the use of linear regression models visualized in scatterplots. Furthermore, various influencing factors, including the model type, the characteristics of the underlying data, and the visual representation of the model outputs, are investigated.

## 3.2 Visual Validation of the Average Value in Scatterplots

The first research on the visual validation of regression models was conducted on the average value (i.e., regression to a constant) in scatterplots. This simplified model is used as a starting point to gain insights into human perceptual validation processes, as there has been no previous research on visual model validation.

The paper was published and presented at the IEEE Visualization conference:

**D. Braun**, A. Suh, R. Chang, M. Gleicher, and T. von Landesberger. Visual validation versus visual estimation: A study on the average value in scatterplots. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 181–185, 2023.
doi: 10.1109/VIS54172.2023.00045

The supplementary material of the paper, including the study data and results, the study documentation, as well as the Python code for the data and stimuli generation, is publicly available at OSF.

I am the primary author of this publication. In this role, I was responsible for the design, implementation, data collection and analysis, as well as the writing and publication of the work. The specific contributions of myself and my co-authors to this publication are outlined below according to the Contributor Roles Taxonomy (CRediT):

**D. Braun**: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Software, Visualization, Writing – Original Draft. *A. Suh*: Writing – review & editing. *R. Chang, M. Gleicher, T. von Landesberger*: Supervision, Conceptualization, Methodology, Writing – review & editing.

# Visual Validation versus Visual Estimation: A Study on the Average Value in Scatterplots

Daniel Braun[1], Ashley Suh[2], Remco Chang[2], Michael Gleicher[3], Tatiana von Landesberger[1]

[1]University of Cologne
[2]Tufts University
[3]University of Wisconsin-Madison

**Abstract:**

We investigate the ability of individuals to visually validate statistical models in terms of their fit to the data. While visual model estimation has been studied extensively, visual model validation remains under-investigated. It is unknown how well people are able to visually validate models, and how their performance compares to visual and computational estimation. As a starting point, we conducted a study across two populations (crowdsourced and volunteers). Participants had to both visually estimate (i.e, draw) and visually validate (i.e., accept or reject) the frequently studied model of averages. Across both populations, the level of accuracy of the models that were considered valid was lower than the accuracy of the estimated models. We find that participants' validation and estimation were unbiased. Moreover, their natural critical point between accepting and rejecting a given mean value is close to the boundary of its 95% confidence interval, indicating that the visually perceived confidence interval corresponds to a common statistical standard. Our work contributes to the understanding of visual model validation and opens new research opportunities.

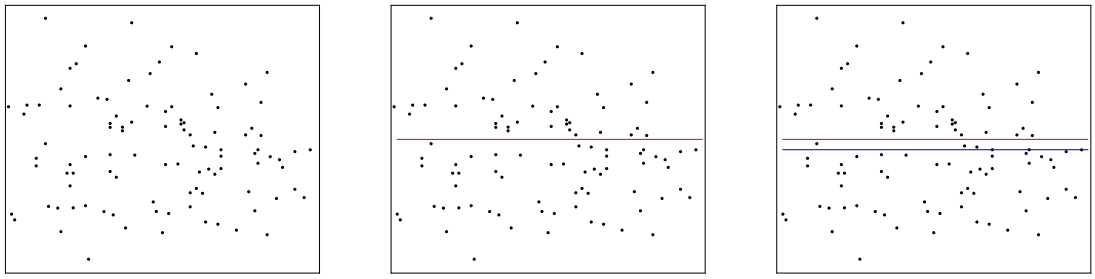
*IEEE Visualization Conference, 2023*

## 1. Introduction

In today's data-driven world, individuals with a range of statistical expertise are tasked with visually validating statistical models fitted to data for a variety of purposes. The general public use visual model validation when consuming news media visualizations, e.g., to become informed about changes of COVID-19 cases over time (such as using a 14-day moving average model visualization [55]). Similarly, domain experts also use visual validation to quickly determine whether a particular model is a good fit to the underlying data (e.g., meteorologists validating a model for assimilation).

To perform model validation, an individual checks that a given model result is a good fit to the data. This model verification can be done computationally or visually; however, statistics computed to validate models are often insufficient to fully describe the underlying data and model's quality [122]. An example is Simpson's paradox, where the evaluation of groups differ depending on whether or not they are divided into subgroups [21, 169]. Thus, it is essential in practice to validate computed models not only through statistical tests, but also visually to ensure that they are accurate and reliable.

Despite the prevalence of *visual model validation* [130], there is markedly little research done to understand viewers' ability to perform these processes. Instead, most research in the perception of statistical modeling has focused on *model estimation* [52, 80, 85, 95, 106, 132, 152, 196, 198, 199], that is, the ability of an individual to perceive, draw, or predict a model that is appropriately fitted to the data. While these studies help inform our understanding of experts and non-experts' strengths and limitations in *estimating* models, they do not advise us of an individual's ability to *validate* how well a statistical model fits



(A) Scatterplot as shown in the visual estimation tasks.

(B) Scatterplot as shown in the visual validation tasks.

(C) Scatterplot with the true average value of the dots.

FIGURE 1: Example scatterplot shown to participants in our user study to investigate the differences between (a) visually estimating and (b) visually validating the average value of the shown data. In (c), the red lines indicate the upper border of the statistical 95% confidence interval (CI) of the average value and the blue line shows the data's true average value.

the data. Thus, it is currently unknown how well people validate models in comparison to their estimation ability.

We seek to establish a baseline for understanding the differences between an individual's ability to perform model estimation and validation. As an initial exploration, we chose to use *averages* (mean expected value) in scatterplots. The model of averages is a simple but generalizable model that is widely used in visual analysis applications [59, 65]. It corresponds to the result of a linear regression to a constant that summarises the underlying data and provides a simplified representation of its central tendency. It also allows us to compare our results against existing literature on visual model estimation [80, 95, 199] and to follow recognized practice of "finding trends" in scatterplots [129].

Using this setting, we compare the visual processes of model validation and estimation and answer the following research questions:

- **RQ1**: Are individuals able to perform visual validation consistently and without bias for averages in scatterplots?
- **RQ2**: How does performance in visual validation relate to the accuracy of visual estimation in scatterplots?

We conducted a human subjects study with two participant groups to address these research questions. Using a between-subjects study design, participants either had to validate whether a shown line in a scatterplot represents the average value of the dots (Figure 1b) or had to estimate the average line on their own (Figure 1a). Our study finds that the participants are consistent and unbiased (i.e. not influenced by data distribution) when deciding the threshold for a model to be accepted as valid. Furthermore, the required level of accuracy for validation was found to be lower than what can be estimated through visual inspection. We find that the critical point between accepting and rejecting a line is close to the boundary of the 95% confidence interval (CI) of the true mean value (i.e., a common statistical criteria for validity). These results have two major implications. First, viewers accept models that are less accurate than they can estimate visually. Second, individuals' ability to judge a model within the 95% CI suggests that the visual perceived confidence interval corresponds to a common statistical standard.

Overall, our work contributes towards a new understanding of how well people can visually validate a model when given data. Our study provides a baseline for contributing towards this understanding, opening up future work for generalization beyond scatterplots and the model of averages.

## 2. Related Work

**Visual Model Estimation** Research on visual model estimation has investigated the concept of ensemble coding, the rapid extraction of visual statistics on distributions of spatial or featural visual information to estimate actual statistics on data [181]. In particular, the estimation of linear regression to a constant (i.e., average) has been extensively studied for scatterplots: Hong et al. explored the influence of a third dimension, encoded by color or size of the dots, on mean estimation [95]. They found that people's mean position estimations are biased towards larger and darker dots. Similarly, Gleicher et al. investigated the perception of average dot height in multi-class scatterplots and found that the perceptual process is robust against more points, sets, or conflicting encodings [80]. Additionally, Yuan et al. found that the visual estimation of averages depends on the visual encoding of the data and that people use primitive perceptual cues for their estimation [199]. In line with these previous works, this paper uses scatterplots and averages as a baseline for visual model validation research.

Visual estimation of linear trends, e.g., correlations of two-dimensional data, has also been considered from several aspects. Rensink and Baldridge examined the statistical properties at which a correlation can be perceived for data in scatterplots [152]. Rather than statistical properties, Yang et al. focused on visual data patterns in scatterplots and their effect on the perception of correlations [198]. Their results suggest that visual features, such as bounding boxes, influence people's correlation judgments. Xiong et al. consider the correlation in scatterplots from a data semantics point of view and found that people estimate the correlation more accurately with generic axis labels than with meaningful labels [196]. Comparisons and rankings of different correlation visualizations based on Weber's law showed measurable differences between various designs (with scatterplots being the best) and that the performances differ significantly for positive and negative correlations [85, 106]. These works support our choice of scatterplots for our study and their findings influenced our hypotheses and stimuli design.

Research has also been conducted on the visual estimation of more complex models. For example, Correll and Heer investigated people's ability to perform "regression by eye" for different types of trends and visualization types [52]. Their results showed that individuals can accurately estimate trends in many standard visualizations; however, both visual features and data features (e.g., outliers) can affect the results. Newburger et al. focused specifically on fitting bell curves to different visualization types [132]. They found that people are accurate at finding the mean, but tend to overestimate the standard deviation. We aim to explore whether the findings from these papers also apply to visual validation for less complex models.

**Visual Model Validation** The literature on visual model validation is currently limited. Correll et al. evaluated scatterplots, histograms and density plots as means to support data quality checks [53]. Their findings suggest that problems arise as soon as overplotting occurs, which informed the design of the scatterplots in our study.

Visual model validation is particularly significant in machine learning. Chatzimparmpas et al. provide an overview of how visualization is currently used to interpret

machine learning models [40]. Mühlbacher and Piringer present a partition-based framework for creating and validating regression models that combines the use visualizations with a relevance measure for ranking features [130]. Our work aims to develop an understanding of visual model validation which can inform the development of future visualization and machine learning systems.

Model validation is closely related to the viewers' trust in these models and their visualizations. In their state-of-the-art report, Chatzimparmpas et al. summarize the importance of using visualizations to increase trust in machine learning models [41]. In addition to machine learning, the relationship between visual design and trust is an important area of research that should be considered in future work on visual validation [56, 66, 123, 142, 154].

## 3. User Study

To address our research questions, we first conducted an exploratory pilot study. Based on the results of the pilot, we formulated hypotheses that were subsequently tested in a confirmatory main study.

### 3.1 Experimental Design

The same experimental design and structure was used for both the pilot study and the main study. We used two **tasks** to compare the following visual processes:

- *Visual estimation*: Participants had to draw a horizontal line in the plot by hovering over the image with the mouse and clicking on the desired position on the y-axis (Figure 1a).
- *Visual validation*: Participants were shown a scatterplot with a line already drawn and were asked to indicate whether the line was "too high," "too low," or "about the same" in relation to their perceived true mean value of the dots (Figure 1b).

We chose a **between-subject design** of our study to prevent learning effects between the tasks and to reduce the number of trials per participant [38]. To maintain consistency, we used the same **plots** for both the validation and estimation trials. Each scatterplot, which had dimensions of $100 \times 100$ pixels, contained 100 data points uniformly distributed on the x-axis. The only difference between the two between-subject groups was the lines shown in the validation tasks.

For the analysis, we defined the deviation of a line based on the confidence interval *(CI)* calculation, quantifying the *deviation* from the true mean as a proportion of the standard deviation:

$$\text{shown value} = \text{true mean} + \textit{deviation} \cdot \text{standard deviation} \tag{1}$$

Using regression calculation, the expected acceptance range for lines falls within the 95% CI. Given our fixed number of data points and distributions, the 95% CI is set at a deviation of 0.198. Thus, all lines with a lower deviation should be considered acceptable.

The **study procedure** began with a training phase, followed by the experimental study. Each page of the study interface displayed one trial (i.e., one plot) and response times were recorded. A display size of at least 13" was recommended. The order of the trials (i.e., the order in which participants saw different deviations) was randomized to minimize learning effects. At the conclusion of the study, participants were asked to describe their strategies for completing the tasks and rate the task's difficulty on a 5-point Likert scale [186].

### 3.2 Pilot Study

To gain initial understanding of visual validation and its differences from visual estimation, we conducted a pilot study with 12 participants (7 for validation and 5 for estimation). Each participant answered 30 trials with pseudo-randomized y-coordinates of the scatterplots. The lines shown in the validation task had random deviations within $\pm[0.0, 1.0]$ from the true mean of the data points.

**Results**: The errors in the estimation tasks were roughly normally distributed with $\mu = 0.02$ and a maximum deviation of 0.7. Lines shown in the validation tasks with deviations within the 95% CI ($dev < 0.198$) had acceptance rates of at least 80%, while every line with $dev > 0.198$ had an acceptance rate of at most 30%. Overall, participants were more accurate in visual estimation than visual validation, and they exhibited a slight difference in judgment between positive and negative deviations in the validation tasks. Most participants reported using a "counting" strategy to solve the tasks, approximating the number of points above and below a given line.

### 3.3 Main Study

Analysis of the pilot study's results led to the hypotheses:

- **H1**: The accuracy of visual validation is lower than the accuracy of the visual estimation when perceiving the mean value of points in a scatterplot.
- **H2**: People's critical point between accepting and rejecting a given mean value is close to the boundary of the 95% CI.
- **H3**: For visual validation, the results differ between positive and negative deviation from the true mean.

#### *3.3.1 Stimuli Design*

Based on the findings from the pilot study and for matching the distribution assumption of linear regression residuals, the y-coordinates of the points were generated from a normal distribution with random mean between 30 and 70 and standard deviation between 15 and 25. Following the literature [80, 95], we focused on the perception of only one dimension (i.e., the y-axis). The adoption of a normal distribution is prevalent across numerous applications (e.g., as a pre-requisite of least square regression) and provides consistent conditions throughout all trials, given that the pilot study results were partially influenced by the stimulus. Thus, the level of trial difficulty was determined by the deviation of the displayed line in the validation tasks. Since most participants approximated the median instead of the mean in the pilot study, we made sure that the number of points above and below the true mean differed by at least 10% to discourage the use of "bounding boxes" [198] as a perceptual proxy.

The pilot study showed that lines with deviations greater than 0.7 were consistently rejected. Thus, we used lines with evenly distributed deviations in the range of $\pm[0.0, 0.7]$ to determine participants' acceptance threshold. Consistent with previous work [198], logistic regression was chosen to analyze the validation task (see Subsubsection 3.3.3). A power analysis of the logistic regression of the pilot study indicated that a sample size of at least 50 was necessary to obtain a meaningful model [128]. Therefore, the study included 50 trials with 25 lines with a positive and 25 lines with a negative deviation in the validation task.

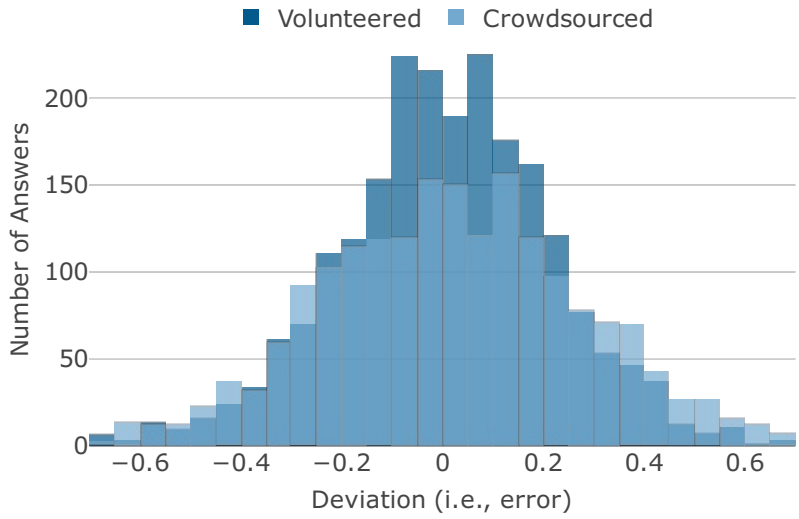



FIGURE 2: Histogram of the deviations of the drawn lines in the estimation tasks for both populations.

### 3.3.2 *Experimental Setting & Participants*

To increase the generalizability of our findings, we conducted an online study with two different populations. The first population consisted of 100 individuals who volunteered to participate after seeing advertisements in university lectures and mailing lists. After filtering out 14 participants who failed attention checks, 42 individuals completed the validation and 44 the estimation task. For the second population, we recruited 90 participants via crowdsourcing platform Prolific [145]. After filtering for attention checks, 42 individuals completed the validation and 40 the estimation task.

In both populations, most participants were between 20 and 30 years old (43% for volunteered, 72% for crowdsourced) and nearly evenly split between women and men (volunteered: 46% F, 50% M, 4% other; crowdsourced: 47% F, 52% M, 1% other). The level of education and experience with statistical model estimation was slightly higher among the volunteers.

### 3.3.3 *Analysis*

In the validation task we measured whether participants accepted the displayed line as the true means. To ensure comparability with the estimation task results, we transformed the responses to binary results. Logistic regression was then applied to the acceptance rates of the shown lines, which is a technique that has been used in previous work [198]. In the estimation task, we measured whether participants were able to draw (estimate) the true means. The estimation errors were measured as the deviation of the lines drawn by the participants.

For statistical testing, we first ran a Shapiro-Wilk test on the given responses and response times to see if they were normally distributed. Although none of the tests were positive, a visual inspection of the acceptance rates and estimation errors suggested that this was due to the large sample size (see Subsubsection 3.3.4). Therefore, we used t-tests to compare these results. We used a Kruskal-Wallis test for the response times and a chi-squared test for comparing Likert responses.

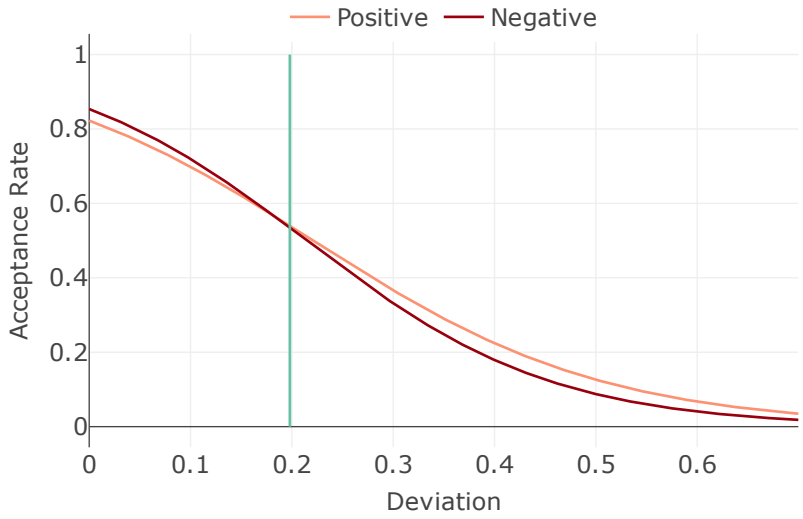


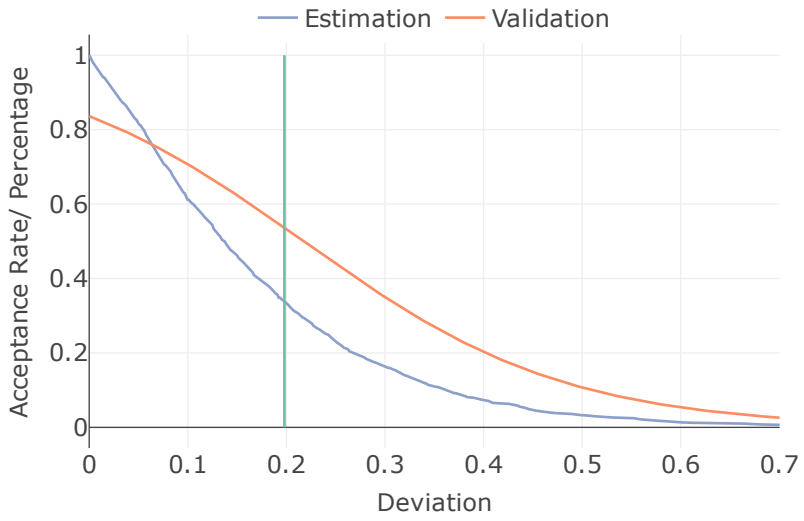

FIGURE 3: Comparison of the logistic regressions for the acceptance rates of positive and negative deviations of the volunteered population. Green line: statistical 95% CI.

### *3.3.4 Results*

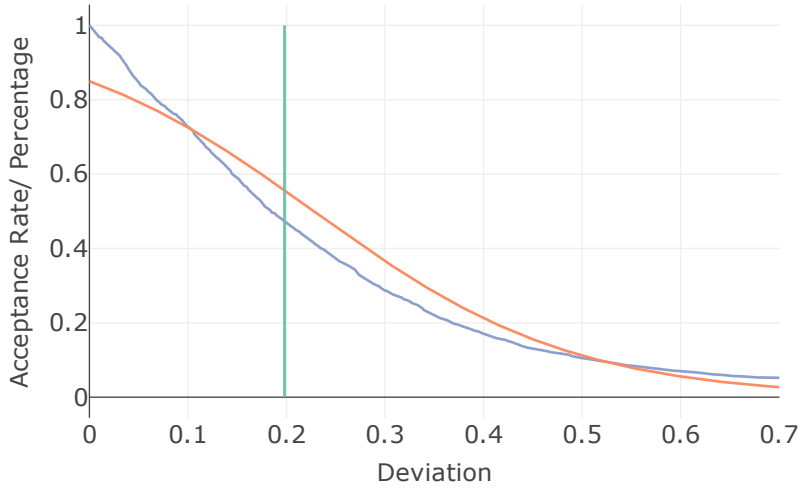
**Distribution and Bias** For both populations, the estimation errors resembled a normal distribution (Figure 2; $\mu_{volun} < 0.01$, $\mu_{crowd} = 0.06$), consistent with the regression assumptions. The same is true for the acceptance rates of the lines displayed for validation ($\mu_{volun} = 0.01$, $\mu_{crowd} < 0.01$). The means of the errors/ acceptance rates being zero indicate that neither of the processes is biased. This finding is supported by a comparison of the acceptance rates for positive and negative deviations from the mean (Figure 3), which showed no significant difference ($p_{volun} = 0.29$, $p_{crowd} = 0.72$). This rejects our hypothesis **H3**.

**Accuracy of Visual Validation vs. Visual Estimation** Since both estimation errors and acceptance rates were approximately normally distributed, we considered positive and negative deviations cumulatively as absolute deviations. We compare the logistic regression for the acceptance rates of the validation task with the cumulative distribution of the estimation errors in Figure 4. For the volunteered population, the acceptance threshold for validation was less accurate than the visually estimated ones ($p < 0.01$, cohensD = 0.35). For the crowdsourced population, the logistic regression for validation was almost identical ($p = 0.51$), but the estimation errors were higher. Although the accuracy of the visually accepted lines was lower than the accuracy of the visually estimated lines, the difference was not significant ($p > 0.29$). In summary, **H1** is partially accepted.

**Critical Point of Validation** Figure 5 shows the raw acceptance rates per deviation and the corresponding logistic regression exemplary for the volunteered population. The inflection points of the logistic regressions correspond by construction to the 50% acceptance rate. Their values $dev_{volun} = 0.217$ and $dev_{crowd} = 0.228$ were very close to the boundary of the statistical 95% CI ($dev = 0.198$). Moreover, the critical points of the individuals' logistic regressions resembled a normal distribution near the 95% CI ($\mu_{volun} = 0.223$, $\mu_{crowd} = 0.241$). This indicates that people's perceived visual confidence interval matches the statistical 95% CI and accepts hypothesis **H2**.




(A) Volunteered population.




(B) Crowdsourced population.

FIGURE 4: Validation and estimation accuracy for both populations (absolute deviation). Blue line: Cumulative distribution for the estimation errors. Orange line: Logistic regression for the validation acceptance. Green line: Statistical 95% CI.

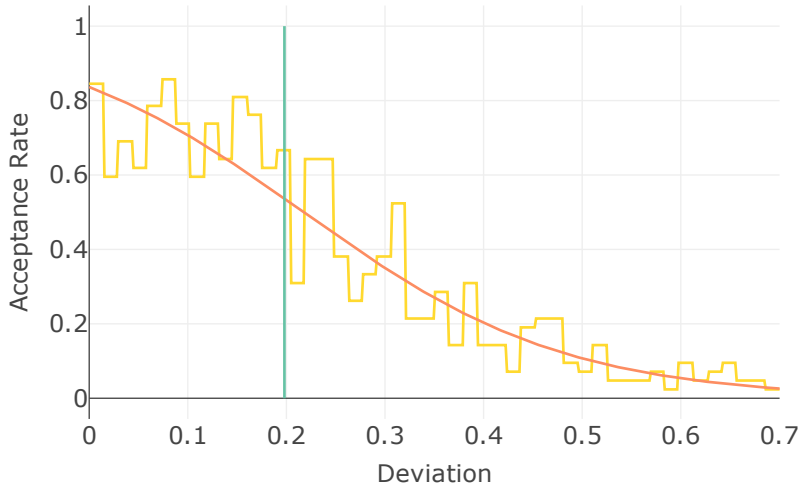



FIGURE 5: Acceptance rate of the shown lines in the validation tasks (absolute deviation) of the volunteered population. Yellow line: The raw percentages per deviation. Orange line: The logistic regression of the yellow line's data. Green line: Statistical 95% CI.

**Difficulty and Response Time** For both populations, no significant difference in response time (*Wilcoxon-test:* volunteered: p-value = 0.94, $\mu_{val}$ = 11.3sec, $\mu_{est}$ = 11.8sec; crowdsourced: p-value = 0.50, $\mu_{val}$ = 10.2sec, $\mu_{est}$ = 10.5sec) and reported difficulty (*chi-squared test:* volunteered: p-value = 0.26, $\mu_{val}$ = 2.76, $\mu_{est}$ = 3.16; crowdsourced: p-value = 0.43, $\mu_{val}$ = 3.14, $\mu_{est}$ = 3.22) was found between the two tasks.

**Self-Reported Strategies** For visual estimation, most participants derived the mean using a perceptual proxy and "adjusted it for outliers" (without having true statistical outliers, the participants probably meant points that were a bit off). These perceptual proxies were based on the density of the scatterplot, the median, or the distance between the highest and lowest points.

In the visual validation tasks, the majority of the participants validated the line based only on the number of points above and below the line. However, some persons judged based on the perception of the density and structure of the points. Six participants "estimated [their own] mean and compared it to the shown line."

## 4. Limitations and Future Work

The participants in this study were non-experts in data visualization or statistics, making them not representative of domain experts. We chose to focus on a specific type of data(-distribution), visualization, and model as a starting point to understand visual model validation. Although our findings are statistically sound, the generalizability of our results regarding these aspects remains to be established. Furthermore, we need to investigate the perceptual mechanisms involved in performing visual model validation, and how they compare to the mechanisms of visual model estimation. By imposing a time limit during the trials, participants would likely stop relying on "counting strategies", but instead adopt other perceptual proxies for solving the tasks. It would also be interesting to explore the influence of visual encoding and data patterns (e.g., outliers, shapes), as well as dimension size and number of points. By doing so, we could derive design guidelines for model visualizations that mitigate perceptual biases in visual validation (e.g., the work by Hong et al. [95]) and provide prescriptive instructions on when and how visualizations should include pre-drawn models fitted to data.

## 5. Conclusion

Our empirical user study with two different populations investigated the difference between visual model estimation and visual model validation by using the average value in scatterplots as a baseline. Our findings suggest that visual model validation accepts models that are less accurate than those that are estimated visually. The acceptance level is similar to the common statistical standard 95% confidence interval. Our study provides valuable insights into how humans process statistical information and identifies limitations and potential aspects for future research.

## Acknowledgments

The authors would like to thank all study participants and the reviewers, whose suggestions helped to improve this paper. This paper is a result of Dagstuhl Seminar 22331 "Visualization and Decision Making Design Under Uncertaint". This work has been partially supported by BMBF WarmWorld Project and Risk-Principe Project. This work has been funded in part by NSF Awards 2007436, 1452977, 1940175, 1939945, 2118201.

## 3.3 Visual Validation of Linear Trends in Scatterplots

This follow-up work on visual model validation increased the model complexity and investigated the differences in humans visual validation and estimation abilities for linear trends in scatterplots. Moreover, it examined the influence of standard designs for visualizing model results on the visual validation task.

The paper is published in IEEE TVCG and was presented at the IEEE Visualization conference:

**D. Braun**, R. Chang, M. Gleicher, and T. von Landesberger. Beware of validation by eye: Visual validation of linear trends in scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):787–797, 2025.
doi: 10.1109/TVCG.2024.3456305

The supplementary material of the paper, including the study data and results, study documentations, as well as the Python code for the data, stimuli, and design generation, is publicly available at OSF.

I am the primary author of this publication. In this role, I was responsible for the design, implementation, data collection and analysis, as well as the writing and publication of the work. The specific contributions of myself and my co-authors to this publication are outlined below according to the Contributor Roles Taxonomy (CRediT):

**D. Braun**: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Software, Visualization, Writing – Original Draft. *R. Chang, M. Gleicher, T. von Landesberger*: Supervision, Conceptualization, Methodology, Writing – review & editing.

# Beware of Validation by Eye: Visual Validation of Linear Trends in Scatterplots

Daniel Braun[1], Remco Chang[2], Michael Gleicher[3], Tatiana von Landesberger[1]

[1]University of Cologne
[2]Tufts University
[3]University of Wisconsin-Madison

**Abstract:**

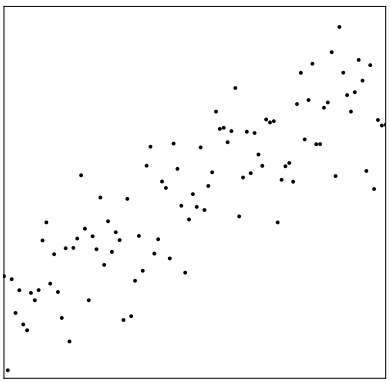
Visual validation of regression models in scatterplots is a common practice for assessing model quality, yet its efficacy remains unquantified. We conducted two empirical experiments to investigate individuals' ability to visually validate linear regression models (linear trends) and to examine the impact of common visualization designs on validation quality. The first experiment showed that the level of accuracy for visual estimation of slope (i.e., fitting a line to data) is higher than for visual validation of slope (i.e., accepting a shown line). Notably, we found bias toward slopes that are "too steep" in both cases. This lead to novel insights that participants naturally assessed regression with orthogonal distances between the points and the line (i.e., ODR regression) rather than the common vertical distances (OLS regression). In the second experiment, we investigated whether incorporating common designs for regression visualization (error lines, bounding boxes, and confidence intervals) would improve visual validation. Even though error lines reduced validation bias, results failed to show the desired improvements in accuracy for any design. Overall, our findings suggest caution in using visual model validation for linear trends in scatterplots.

*IEEE Transactions on Visualization and Computer Graphics, 2025*

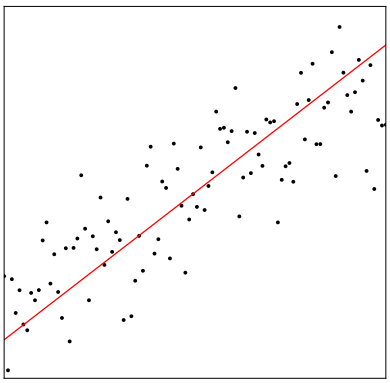
## 1. Introduction

Visual validation of statistical models serves as an important task in modern statistics and machine learning applications. The complexity of models often demands visual inspection for assessing their correctness and reliability [40, 41]. This is crucial, as the model outcomes have great implications in critical domains [96], e.g., the estimation of pandemic outbreaks [55] and meteorological forecasting [102]. Without visualization, traditional statistical metrics are often insufficient in describing the underlying data and model. For example, data sets with significantly different characteristics can share the same numerical metrics [21, 122, 181]. As a result, visualization researchers have advocated for visual validation of statistical models as a core part of data analysis.

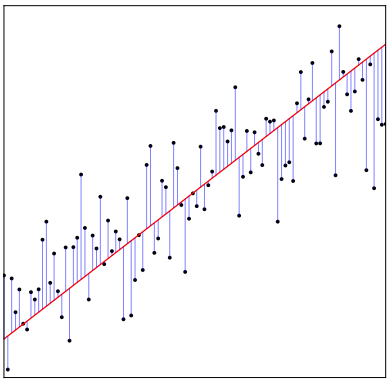
To date, most research on the perception of statistical models has focused on *visual estimation* – individuals' ability to visually fit a model to data [46, 52, 80, 85, 95, 106, 126, 132, 136, 151, 152, 176, 177, 196, 198, 199]. While these studies contribute to our understanding of the visual estimation process, there is a lack of research on *visual model validation* – individuals' ability to assess the fit of a given model to the underlying data. In a recent study by Braun et al. [30], the authors found significant differences in individuals' performance in visual validation and estimation of averages in scatterplots. In this paper, we build upon this work and investigate the accuracy and effectiveness of visual model validation for a more complex model – *linear trends* in scatterplots. While the average value offers a fundamental measure in one dimensions, the exploration of linear trends provides valuable insights into the relationships between two data dimensions, which can



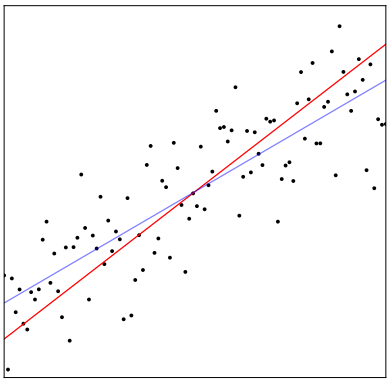
(A) Scatterplot to visually estimate linear trend.



(B) Trend line to be visually validated.



(C) Visual design of adding error lines of ordinary least squares regression.



(D) Example stimuli construction: True regression line (blue) and line to be validated (red, see Figure 1b).

FIGURE 1: Example scatterplot shown to participants in our user studies to investigate the visual validation of linear trends. In experiment 1 we tested the difference between visual validation (b) and visual estimation (a), while in experiment 2 (c) we compared different visual designs for regression to support visual validation. In (d), the blue line shows the true OLS regression.

be particularly relevant in machine learning applications where linearity assumptions are common. Scatterplots are the standard approach for showing relationships between two quantitative variables [129].

Linear trends can be interpreted in two different ways [20]: *Regression* and *correlation*. Regression estimates the underlying linear relationship between two variables to find a function that predicts the value of one variable based on the other. Correlation measures the strength of the linear relationship between two variables. The closer points lie to a straight line, the stronger their relationship. Due to the relevance of regression models for decision making in wide-ranging fields [57, 58, 127, 178], visual validation of regression models is important. In this paper, we investigate the perception of linear regression models and answer the following research questions:

- **RQ1**: How does performance in visual validation of linear trends relate to the accuracy of visual estimation?
- **RQ2**: Can common visual designs enhance the performance of visual validation of linear trends?

For RQ1, we examined individuals' perception of the slope of a linear trend in a between-subject user study. Participants were randomly assigned to either validate the slope of a shown trend line in a scatterplot (Figure 1b) or estimate the trend line on their own (Figure 1a).

Our results confirm the previous findings of Braun et al. [30] that participants are more accurate in model estimation than in model validation. Further, we found that participants systematically overestimated the trends' slope and were particularly inaccurate at recognizing trend lines with a high slope value (i.e., lines that are "too steep"). This implies that the perception of the trend lines is biased – not symmetrical, i.e., lines with slopes that are too high and too low are perceived differently. A post-hoc analysis revealed that participants' responses were more consistent with the non-standard orthogonal regression (ODR: minimizes orthogonal distances between data points and regression line) than with the usual vertical regression (ordinary least squares - OLS: minimizes vertical distances between data points and regression line), since ODR regression has a higher slope per calculation (see Figure 6). This implies they assumed errors in both x and y variables rather than just one. Our results show similar effects between positive and negative trends.

With respect to visual design, several recent studies have found significant effects of visualization on model estimation [85, 126, 176, 177, 181, 198]. We similarly investigate whether the addition of augmentations to the visual designs improve individuals' performance in the visual validation of models (RQ2). To answer this research question, we modified the visual designs and repeated our first study on the slope validation. Using the same data, we added three visual designs to the shown trend line - OLS error lines, 95% confidence intervals, and bounding boxes (see Figure 12). The results showed that the addition of error lines reduced the bias in recognizing lines with slopes that are either too high or too low. The addition of bounding boxes slightly increased visual model validation accuracy. However, none of the designs lead to the desired improvements (i.e, higher acceptance rate for correct models and higher rejection rate for incorrect models). In fact, participants reported an increased task difficulty with the addition of visual designs with no benefits to task completion time.

Altogether, the results of our two studies find evidence to caution when using the common practice of "validation by eye" [96] – visually validating statistical models by overlaying a visualization of the model over raw data. Further research is needed on how to support individuals in the visual validation process.

The paper is structured as follows: After an overview of current research studies on visual validation and visual estimation in Section 2, we give details on the general data, study, and analysis design in Section 3. The two experiments - validation versus estimation and visual designs- as well as their analysis and results are presented in Section 4 and Section 5. Section 6 discusses limitations of our work and possible future work.

## 2. Related Work

**Visual Model Validation** The research topic of visual model validation has received little attention in the past, but has become more prominent recently as it is recognized as an essential part of exploratory data analysis [96]. Braun et al. [30] were the first to investigate the perceptual differences between visual validation and visual estimation using the example of average values in scatterplots. They found that participants were more accurate in estimation than in validation, that the visual validation of averages is unbiased, and that the critical point between accepting and rejecting a given value is

close to the statistical 95% confidence interval. This study motivated this work on the validation of linear trends.

Few other studies relate to our experiment. Majumder et al. [119] examine the visual validation of statistical inference in linear models, in which participants had to visually find the most deviating model (e.g. highest slope) in a small multiple setting. Their study showed that visual tests have higher power than conventional tests when the effect size is large. The findings of Correll et al. [53] on visual validation of data distributions in scatterplots, histograms, and density plots suggest problems with overplotting, which informed the stimuli design in our study.

Most of the time, the possibility for visual model validation is given in an interactive way as part of a visual analytics or machine learning system. Chatzimparmpas et al. [40] give an overview of the current use of visualization for machine learning model interpretation. Bögl et al. [22] provide a visual analytics process to assist domain experts in selecting suitable models in time-series analysis. While the Visual (dis)Confirmation tool by Choi et al. [43] allows users to perform data analysis by automatically generating appropriate visualizations based on hypotheses framed in natural language, Chegini et al. [42] support users in identifying local patterns in large scatterplot spaces by automatically comparing local regions using a model-based pattern descriptor. Mühlbacher and Piringer [130] introduce a partition-based framework for validating regression models both qualitatively via visualizations and quantitatively via a relevance measure for ranking features. Our work is synergistic with the prior work in that it contributes to a better understanding of visual model validation that can improve the development of visual analytics and machine learning systems in the future.

An additional component of model validation is the viewer's trust in these models and their visualizations, which has been widely studied in the past [56, 66, 123, 142, 154]. The survey by Chatzimparmpas et al. [41] summarizes the importance of this relationship between visual design and trust in machine learning. In our study, we tried to minimize the influence of trust by giving the participants as less information as possible without impairing their understanding of the tasks.

**Visual Model Estimation** Most studies on the perception of statistical models aimed to understand visual estimation. In addition to research on the estimation of average values [80, 95, 199], there are many papers on the estimation of linear trends.

Correll and Heer [52] examined the basic perceptual process of visual estimation and had participants perform "regression by eye" for linear trends in scatterplots and other visualizations. They find that an individual's ability to estimate the slope of a linear trend with respect to the least squares regression model depends on both visual features and data features, without bias for positive and negative trends. Ciccione and Dehaene [46] were also interested in the accuracy and bias of visual regression estimation in scatterplots. Their results indicate that people consistently overestimate trends. These works played a key role in our study, hypotheses, and stimuli design.

A different aspect of linear trends is the correlation of the two data dimensions. Rensink and Baldridge [152] investigated the influence of statistical properties on the perception of correlation in scatterplots. Besides statistical properties, correlation perception research dealt with the influence of visual designs, features, and ensemble coding [181]. Yang et al. [198] showed that visual features, such as bounding boxes, are used as proxies for estimating correlation in scatterplots. Xiong et al. [196] found that people estimate correlations more accurately in scatterplots with generic axis labels than with semantic labels. Comparisons of different correlation visualizations based on Weber's law ranked

scatterplot as the best visualization design and showed that performances of correlation estimations differ for positive and negative correlations [85, 106]. These studies confirm our choice of scatterplots as the visual design for our studies, and their results have implications for our stimuli design.

## 3. Experimental Design

We conducted two experiments to gain insights into the perceptual process of the visual validation of linear trends and to answer our research questions:

**RQ1**: How does performance in visual validation of linear trends relate to the accuracy of visual estimation? and

**RQ2**: Can common visual designs enhance the performance of visual validation of linear trends?

The **first experiment** aimed to answer RQ1. Therefore, it contained two *tasks* to compare the following perceptional processes:

- *Visual validation*: Participants were shown scatterplots with an already drawn trend line (see Figure 1b). They were asked to indicate whether the line was "too steep',' "too flat", or "about the same" (i.e., the shown line represents the actual trend) in relation to the true slope of the linear trend of the data.
- *Visual estimation*: Participants were asked to fit trend lines to the data in the scatterplots (see Figure 1a) by adjusting the slope of the line by moving a slider.

The **second experiment** aimed to answer RQ2 by evaluating three regression visualization designs for validation. In this study, participants only had to perform the validation task described in experiment 1.

### 3.1 Study procedure

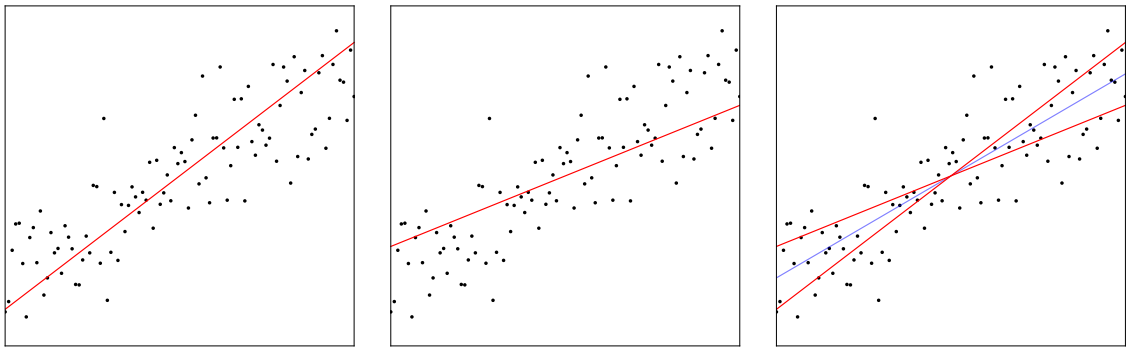
The two experiments addressed different research questions, but used the same study structure and data.

The *study procedure* began with demographic questions about the participants, followed by a short training period for the participants to familiarize themselves with the study interface. To avoid bias in the participants' responses, we did not provide training feedback. To minimize learning effects, the order of trials was randomized. In the study interface, each page displayed one trial (i.e., on plot). For further analysis, response times were recorded. At the end, we asked the participants for their strategy in performing the task and to rate the difficulty of the tasks on a 5-point Likert scale (1 – very difficult, 2 – difficult, 3 – neutral, 4 – easy, 5 – very easy) [186].

Using a *between-subject study design* (i.e. the participants were divided into two groups and had to either estimate or validate trends), we prevented learning effects between the trials and reduced the number of trials per participant [38]. To ensure consistency and comparability, the same data were used in both studies for all of the between-subject groups. The only difference between the two experiments was the visual representation of the data in the scatterplots (see Figure 1b and Figure 1c as an example).

### 3.2 Data Generation and Stimuli Design

Our data generation is inspired by the approaches used by Braun et al. [30] and Correll and Heer [52]. Each trial displayed a scatterplot with a size of $700 \times 700$ pixels. Therefore, we recommended a screen size of 13" or larger. The scatterplots contained 100 data points in the range of $[0, 1]x[0, 1]$ uniformly distributed along the x-axis. We used the standard regression model for the point generation in order to be able to compare



(A) Positive slope deviation of 0.5. (B) Negative slope deviation of 0.5. (C) True regression line (blue) and the two deviating lines from (a) and (b) (red).

FIGURE 2: Example of the same amount of positive and negative slope deviation in the validation task.

the validation and estimation results with the ordinary least squares (OLS) regression: $y = ax + b$. We used this function to generate set of points along particular trends. The trend lines were centered in the scatterplot (as by Correll and Heer [52]), i.e., $f(0.5) = 0.5$, and both positive and negative slopes were used. The y-coordinates of the resulting data points were then permuted using a normal distribution. Investigating the perception of the slope of linear regressions, the slope parameter $a$ was the only variable in our studies. Centering the target trends in the plot ensured that participants could always estimate the true regression line by solely manipulating the slope value.

To keep the difficulty level as similar as possible between trials, the standard deviation of the normal distribution was fixed to 0.1 and the slopes of the trends were set to a range of $[0.35, 0.65]$. Since the permutation of the y-coordinates could lead to a deviation of the resulting regression from the original target trend, we performed rejection sampling to ensure that the slope of the trend of the resulting points was within $10^{-3}$ of the target slope.

To be able to measure the accuracy of visual validation, we showed the participants trend lines with slopes deviating from the true regression slope. Due to the centering of the underlying trends, a variation of the slope means a rotation of the line around the point $(0.5, 0.5)$. We define the **deviation** of the displayed line as the deviation from the true slope as a proportion of the regression's standard error:

$$\text{shown slope} = \text{true slope} + \textit{deviation} \cdot \text{standard error} \tag{1}$$

Figure 2 shows an example validation trial with the same amount of positive and negative slope deviation (0.5 in this example). The standard error of the regression is data dependent and represents the average distance between the points and the regression line. This allows us to compare and analyze trials with different data sets regardless of their properties, since an absolute change in slope is perceived differently in graphs with high and low point dispersion.

The deviation definition is based on the calculation of the regression confidence interval *(CI)*. In statistical analysis, the 95% CI is a common measure for the uncertainty in models [9]. By construction – in our setting with a fixed number of points and distribution – the slope's 95% CI is set at a deviation of 0.198. We use the confidence interval to compare the user study results with statistical quality measures. In a statistical sense, all

lines with a smaller slope deviation should be considered acceptable. However, people may have a smaller confidence interval and reject lines with deviations less than 0.198.

Based on the result of the study for validation and estimation of the average value [30], where deviations greater than 0.7 were consistently rejected, we used the same deviation range of $[-0.7, 0.7]$ (i.e., we showed lines with a maximum slope deviation of 0.7 based on Equation 1). The deviations we used for the lines shown in the studies were evenly distributed within this range in 0.05 increments (including deviation 0), resulting in a total of **58 trials** (i.e., 58 different data sets) per participant (29 trends with a positive slope, 29 with a negative slope).

In line with the literature [30, 52, 196], we kept the displayed scatterplots as clean as possible to minimize visual distraction by omitting axes marks and labels (see example stimuli in Figure 1). The used colors and marker sizes were the same as in the study by Braun et al. [30].

### 3.3 Analysis Procedure

In order to assess our results in relation to literature, our analysis is similar to that of Braun et al. [30]. For comparability of the estimation and validation results, we transformed a participant's validation responses to binary results: *1* for *accepting* (i.e., "about the same") the shown line, *0* for *rejecting* (i.e., "too steep" or "too flat") the shown line. Similar to previous studies, logistic regression was then applied to the acceptance rates [30, 198] to assess validation accuracy.

The estimation errors (i.e., the deviation in slope of the self-adjusted trend lines) were calculated using the same deviation definition as for validation (Equation 1). This allows us to compare the logistic regression of validation acceptance rates with the cumulative distribution (CDF) of estimation errors.

For statistical testing, we used a multi-stage approach with the standard significant level $\alpha = 0.05$ in all tests. First, we performed a Shapiro-Wilk test on the given responses and response times to test the data for normality, with the result that none of the data fulfilled this property. Based on this, we then used the non-parametric Kolmogorov-Smirnov *(KS)* test for the difference in validation and estimation results (i.e., comparing the logistic regressions of the validation acceptance rates and the CDF of the estimation errors). A Wilcoxon test was uses to test the response times and estimation errors on difference in means. For comparing the Likert responses, we used chi-squared test. For experiment 2 we first performed a Kruskal-Wallis test on the response times and a chi-squared test on the categorical responses to test the visual designs for significant differences. As post-hoc pairwise analysis of the respective tests to compare the different design combinations.

## 4. Experiment 1: Visual Validation versus Visual Estimation

Experiment 1 analyzed the performance in visual validation of linear trends in relation to visual estimation. Based on previous research on model perception [30, 46, 52, 144] and our own assessments when generating the data, we propose the following hypotheses for experiment 1:

- **H1**: *The accuracy of visual validation is lower than the accuracy of visual estimation when perceiving the slope of a linear trend in a scatterplot.*
  The study by Braun et al. [30] for the average value showed that visual estimation provides a higher level of accuracy compared to visual validation. We expect this result to hold for linear trends.

- **H2**: *People's critical point between accepting and rejecting a given trend line when validating is close to the boundary of the 95% CI.*
  Also based on the results of Braun et al. [30] we expect people's "visual confidence" to match the statistical CI, as is true for the visual validation of average values.
- **H3**: *For visual validation, the results don't differ between positive and negative trends.*
  Correll and Heer [52] found no bias in visual estimation of linear trends. We expect visual validation and estimation to be similar in this regard.
- **H4**: *For visual estimation, people overestimate the slope of linear trends.*
  We expect this behavior found in previous research [46, 52, 144] to be the same in our study.
- **H5**: *Perceived task difficulty and task completion time are lower for visual validation than for visual estimation.*
  The validation task is simply a matter of acceptance, whereas the estimation task requires participants to fit a line to the data. Therefore, we expect participants to perceive the validation task as easier and complete it faster.

### 4.1 Experimental Setting and Participants

We conducted an online study on Limesurvey [115] and recruited participants from the crowdsourcing platform Prolific [145]. A total of 122 participants took part in the study. They had to speak English fluently. No restrictions were made for country of residence. We removed the data from 12 participants because of their incorrect answers to the attention question. Out of the 110 remaining participants, 46 completed the validation and 64 answered the estimation task. Most of them were between 20 and 40 years old (86%) and the gender distribution was close to even (F: 48%, M: 49%, other: 3%). Participants' educational levels ranged from high school diplomas to doctorate degrees, with the majority having a bachelor's degree (40%). The overall self-reported expertise in statistical model estimation was relatively low, with 88% of the participants indicating an expertise between 1 and 3 on a 5-point Likert scale. The average time to complete the study was 20 minutes. Participants were compensated with £4.45.

### 4.2 Results

Given our data generation approach based on the ordinary least squares (OLS) regression (see Subsection 3.2), we analyze the responses of the study participants based on this model.

#### *4.2.1 Accuracy of Visual Validation vs. Visual Estimation*

We evaluate the accuracy of visual validation and estimation in relation to the "statistical accuracy" of OLS regression. In statistical terms, all trend lines within the 95% confidence interval are considered to be valid. This means that, in a perfect world, participants should accept all trend lines with a slope deviation of less than 0.198 (i.e., the CI slope deviation) and reject all trend lines with a higher slope deviation in the validation task. In the estimation task, participants should ideally only estimate lines with slope deviations less than 0.198.

To be able to compare validation and estimation results, we summarize the acceptance rates and estimation errors. We combine the results for positive and negative trends and positive and negative slope deviations as absolute acceptance rates and errors. Figure 3 shows the resulting logistic regression for the validation acceptance rates and the cumulative distribution for the estimation errors.

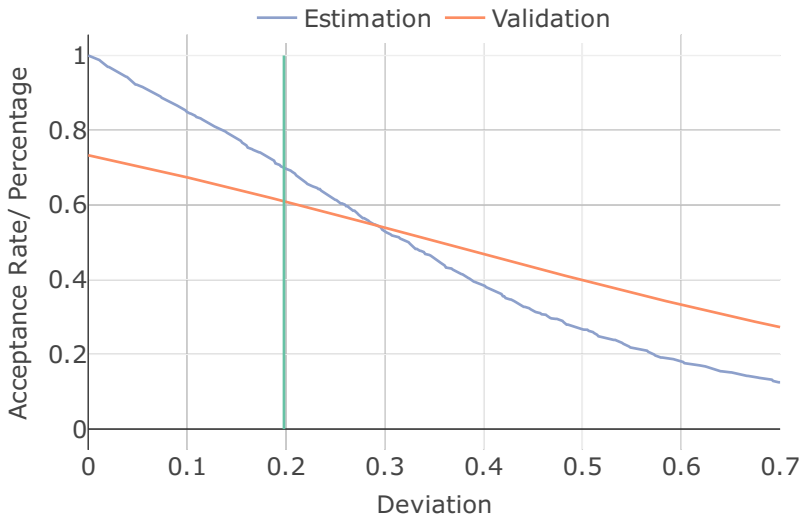


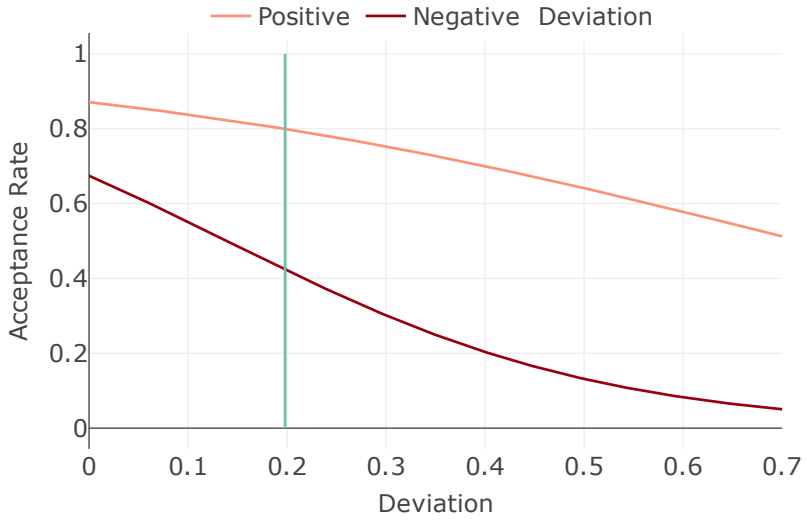

FIGURE 3: Comparison of validation and estimation accuracy (absolute deviation) with respect to OLS regression. Blue line: Cumulative distribution (CDF) for the estimation errors. Orange line: Logistic regression for the validation acceptance. Green line: Statistical 95% CI. Notice that more statistically valid lines were estimated than accepted by validation and more invalid lines were accepted than estimated.

As shown, when the slope deviation is low (e.g., less than the 0.198 – the green line), participants were more accurate in estimating than validating trend lines. For example, about 70% of the participants were able to estimate (i.e., drew) trend lines with slope values below the deviation of 0.198. In contrast, only 60% of the participants correctly validated (i.e., accepted) shown trend lines with the same slope values. When the slope deviation is high, again we see that participants were more accurate in estimating than validating trend lines. For example, only about 20% of the participants estimated (i.e., drew) trend lines with slope values above the deviation of 0.6. In contrast, about 35% of the participants falsely validated (i.e., accepted) shown trend lines with the same slope values. Only in the deviation range adjacent to the confidence interval (0.2 to 0.3) validation is more accurate than estimation. In this statistically invalid range, a slightly higher percentage of lines were estimated than accepted. Notably, the validation acceptance of linear trends has an almost linear relationship with the slope deviation. Ideally, it should have high values at low slope deviations with sharp drop at the 95% CI border (i.e., 0.198).

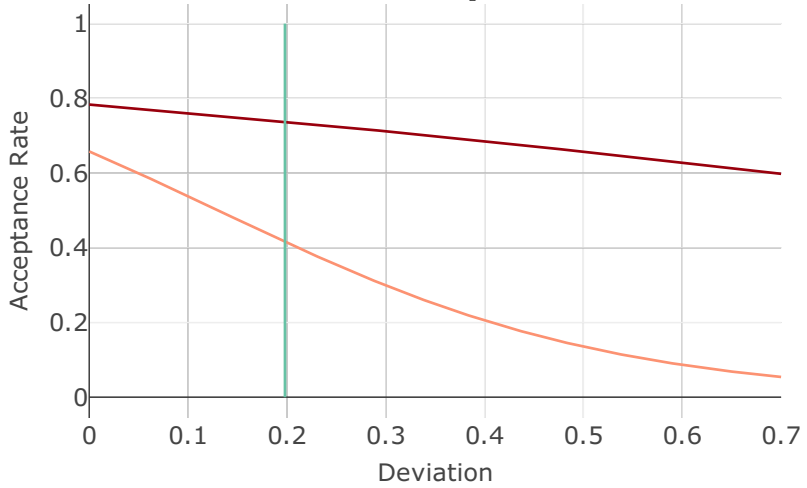
The KS-test showed the two curves to be significantly different ($p < 0.01$, $D = 0.296$). Therefore, **H1** is supported for OLS regression.

The critical points of validation and estimation are precisely these slope deviations, with a 50/50 chance that a line will be accepted or rejected, or estimated with a lower or higher slope. This critical point is slightly lower for estimation ($crit_{val} \approx 0.36 > crit_{est} \approx 0.32$). Moreover, the critical point of validation is much greater than the deviation of the 95% CI (0.198), which does not support our hypothesis **H2** for the OLS regression. The critical points describe an experimental human threshold for the two tasks. People estimated lines that would be correct with the statistical 99.8% CI. For visual validation, they even accepted lines with a deviation greater than the statistical 99.9% CI.

When analyzing the individual performance of the participants in visual validation, it is noticeable that 72% of the participants had an individual critical point greater than the 95% CI and 13% accepted incorrect models more often than correct ones.




(A) Positive slope.




(B) Negative slope.

FIGURE 4: Comparison of the *validation* acceptance rates of positive and negative deviations for positive and negative trends with respect to OLS regression. Green line: statistical 95% CI.

### 4.2.2 *Bias in Positive and Negative Slope and Deviation*

Figure 4 differentiates the validation acceptance rates by positive and negative trends as well as positive and negative slope deviation. For positive trends, lines with positive slope deviations (i.e., trend lines that were "too steep") were accepted significantly more often than lines with negative deviations (see Figure 4a) (KS: $p < 0.01$, $D = 0.814$). For negative trends, the results are mirrored, i.e., lines with a negative slope deviation were accepted significantly more often (Figure 4b) (KS: $p < 0.01$, $D = 0.926$). For both trend directions, lines that were "too steep" were still accepted more than 50% of the time, even with the largest slope deviation. Analysis of individual participant acceptance rates showed that the difference between positive and negative slope deviations was significant for each individual participant. Comparing the logistic regressions of the acceptance rates between positive and negative trends, there is no significant difference for "too flat" lines (KS: $p > 0.99$, $D = 0.022$), but a significant difference for "too steep" lines (KS: $p < 0.01$, $D = 0.336$), indicating a slightly more accurate validation of positive trends. Thus, hypothesis **H3** cannot be rejected.

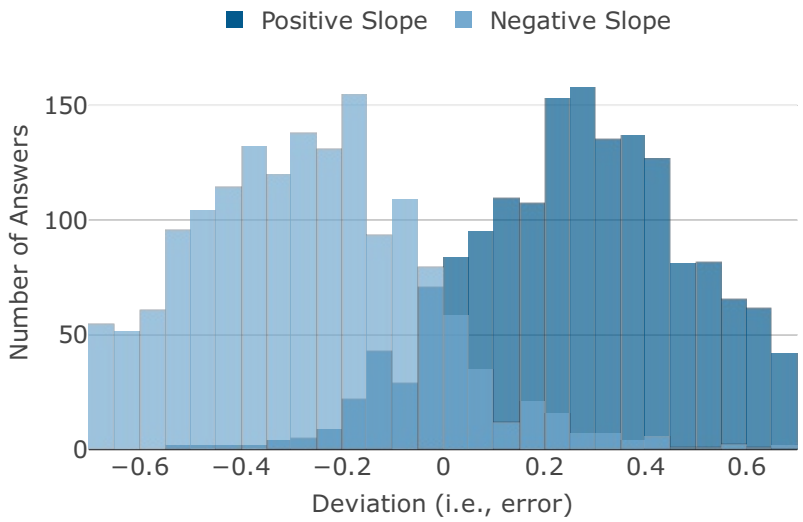



FIGURE 5: Histogram of the deviations of the *estimated* lines for positive and negative trends with respect to OLS regression.

A similar pattern can be observed for the estimation errors (see Figure 5), supporting the results of previous studies [46, 52, 144] and our hypothesis **H4**: People overestimate the slope of trend lines. Without over-estimation, both distributions should be centered at 0 slope deviation. With an average slope deviation of 0.357 for positive trends and −0.359 for negative trends, participants consistently drew the trend lines too steeply (Wilcoxon: $p_{\text{positive}} < 0.01$, $V_{\text{positive}} = 1660685$; $p_{\text{negative}} < 0.01$, $V_{\text{negative}} = 78910$). No significant differences in estimation errors between positive and negative trends could be found (Wilcoxon: $p > 0.52$, $W = 1743165$).

In summary, both perceptional processes – visual validation as well as visual estimation – are biased toward "too steep" slopes for positive as well as negative trends.

#### *4.2.3 Post-Hoc Analysis: OLS vs. ODR*

The results opened a new question: What is the reason why people perform poorly at perceiving trend lines that are "too steep"?

A clue to this question was found in a study by Ciccione and Dehaene [46]. Inspired by their results, we hypothesize that individuals perceive orthogonal distance (ODR) instead of orthogonal least squares (OLS) regression, since ODR regression has a higher slope per calculation. Figure 6 illustrates the difference between the two regression models. OLS regression minimizes the sum of the squares of the vertical distances of the points to the line and assumes noise only in the dependent variable (y-axis). ODR regression minimizes the sum of the squares of the orthogonal distances of the points to the line and assumes noise in both variables. The slopes of the ODR regression lines differ from those of the OLS regression. Thus, taking the ODR regression model into account changes the true slope value, which influences the measured accuracy of visual validation and estimation. We analyzed and compared the results of both tasks with respect to both regression models – OLS and ODR – as a post-hoc analysis to see if it explained the bias in slope perception.

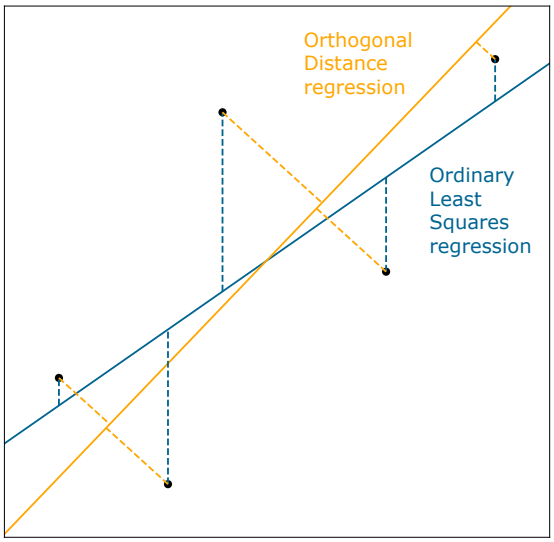



FIGURE 6: Illustration of the difference between ordinary least squares (OLS, blue line) and orthogonal distance (ODR, yellow line) regression, adapted from [46].

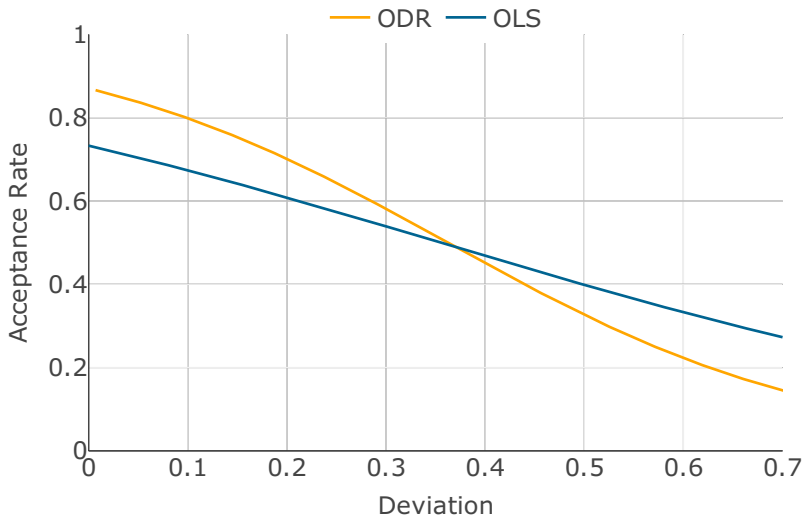



FIGURE 7: Comparison of *validation* accuracy (absolute deviation) with respect to OLS and ODR regression.

**Visual Model Validation:** The acceptance rates for the shown trend lines in the validation task compared for ODR and OLS are visualized in Figure 7. They differ significantly for both models (KS: $p < 0.01$, $D = 0.342$). For ODR regression, more lines with small slope deviation were accepted and more lines with large deviation were rejected. The logistic regressions intersect closely at the 50% acceptance rate, i.e. the critical point between acceptance and rejection of a line remained almost identical for both regression models.

**Visual Model Estimation:** There is also evidence of an improvement in visual estimation (see Figure 8). The estimation errors have improved for both positive and negative trends and the slope is less strongly overestimated compared to OLS ($\mu_{pos} = 0.210$, $\mu_{neg} = -0.216$).

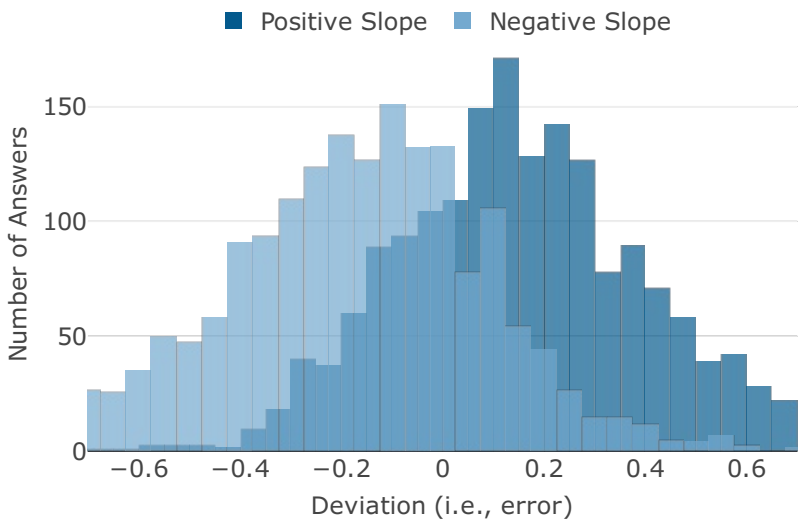



FIGURE 8: Histogram of the deviations of the *estimated* lines for positive and negative trends with respect to ODR regression.

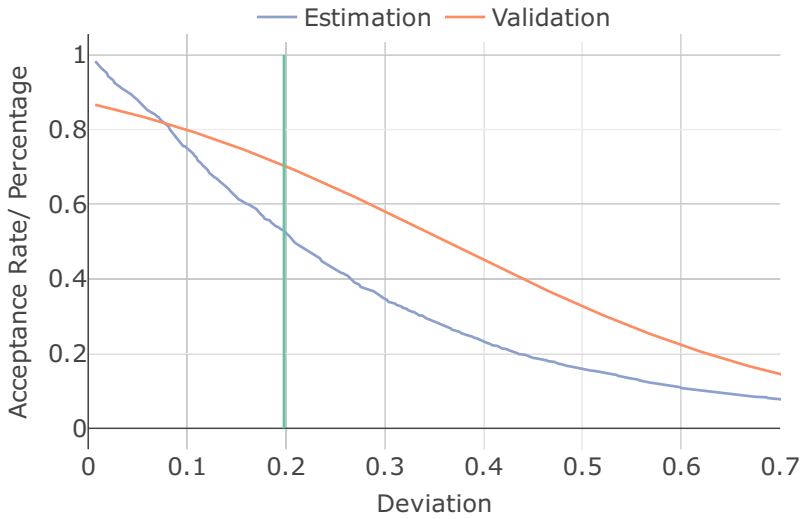
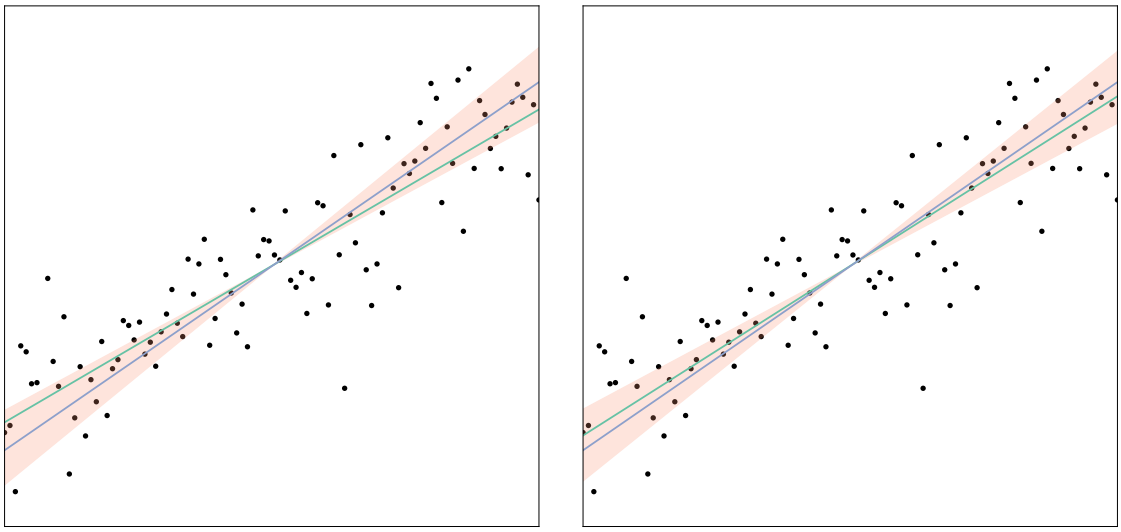



FIGURE 9: Comparison of validation and estimation accuracy (absolute deviation) with respect to ODR regression. Blue line: Cumulative distribution (CDF) for the estimation errors. Orange line: Logistic regression for the validation acceptance. Green line: Statistical 95% CI.

Comparing the two tasks, the differences between estimation and validation were greater for ODR than for OLS (see Figure 9). The critical point of estimation moved closer to the border of the 95% CI for ODR ($crit_{est}$ = 0.208). As a result, the estimation errors were significantly lower than the acceptance threshold for validation (KS: $p < 0.01$, $D = 0.234$).



(A) Mean estimation and validation responses with respect to OLS regression.

(B) Mean estimation and validation responses with respect to ODR regression.

FIGURE 10: "Visual summary" of the results for experiment 1 in an example stimulus. The figures show the true regression line (green) for OLS (a) and ODR (b) together with participants' average response for estimation (blue) and the range of lines with an acceptance rate of 50% or higher for validation (orange). Notice that in figure (b), the blue and the green lines are closer to each other than in figure (a) and the orange range better encapsulate the green line. These suggest that the ODR model better fits participants' perception of trend lines.

Figure 10 illustrates the accuracy results of experiment 1. It visualizes an example stimuli with the true OLS and ODR regression lines together with the average responses for validation and estimation. For estimation, this means a line (shown in blue) with the average deviation of the estimated lines. For validation, this is represented as the range of lines (shown in orange) that would be accepted at least 50% of the time.

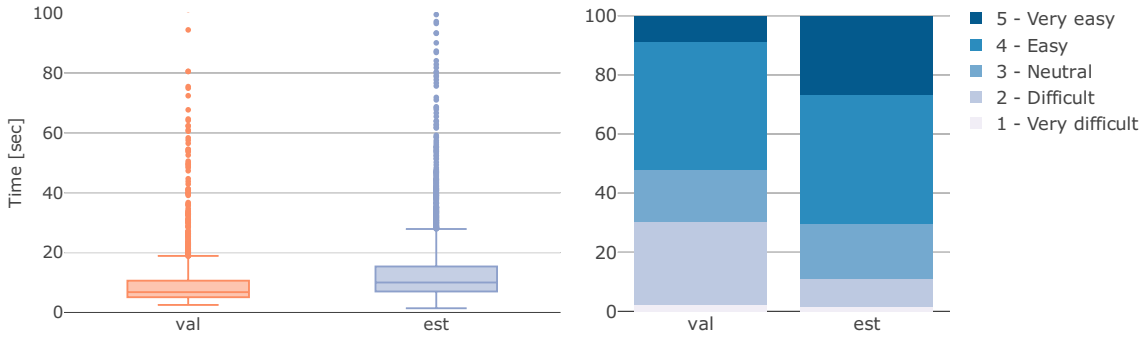
In sum, we found that participants' trend perception was more consistent with the ODR regression model than with the OLS model for both estimation and validation tasks.

#### *4.2.4 Response Time and Difficulty*

Figure 11 shows the response times and the distribution of the Likert scale responses for the task's difficulty. As shown, the response times for visual validation were significantly faster than for visual estimation (Wilcoxon: $p < 0.01$, cohensD = 0.12, $\mu_{val} = 10.85sec$, $\mu_{est} = 13.73sec$). In contrast, participants' self-reported task difficulty was significantly lower for the estimation than for the validation task (Chi-squared: $p < 0.05$, $\chi^2 = 10.08$; $\mu_{val} = 3.28$, $\mu_{est} = 3.84$). Therefore, Hypothesis **H5** cannot be rejected.

#### *4.2.5 Self-Reported Strategies*

Participants' self-reported strategies were provided in free-text form and subsequently summarized by us. For *validation*, the most often strategy was the comparison of the shown line shown with a self-estimated line (responded by 8 participants). Alternative strategies include a counting strategy ($n = 7$) where the participants counted the dots on both sides of the line, and a strategy that references the overall visual image of the visualization ($n = 6$) where the participants checked whether the trend line passed through the center of the area of dots. The latter strategy is similar to the strategy of a "bounding box"




(A) Response times. (B) Self-reported task difficulty.

FIGURE 11: Response times and self-reported difficulty for experiment 1.

around the scatter dots, which was investigated by Yang et al. [198] for the perception of correlation (see Subsection 5.1).

For *estimation*, one strategy was used very often. 14 participants reported using the counting strategy and balancing the number of points on each side of the line. Only two other strategies were mentioned more than once: drawing a line through the middle of the dots ($n = 3$) and mentally connecting the dots to a line chart ($n = 2$).

For both tasks, the comparison of the results per strategy did not show any significant differences.

### 4.3 Discussion

Comparing the perceptual tasks of visual validation and estimation, individuals are more accurate at estimating regression models themselves than validating existing ones. One possible reason for this is that, as noted in the self-reported strategies, when validating a trend line, many people compare the line shown to a self-estimated line. Due to their uncertainty in their own estimation, they then may give themselves a margin of tolerance, which is reflected in the larger accepted deviations for the validated lines. This may indicate that people go through a two-step process during visual validation, with estimation as the first step. However, the response times and the self-reported task difficulty provide contradictory supporting evidence. Although participants reported estimation to be easier than validation, they also needed more time to complete the task. One possible reason for the longer response time for estimation is the interactivity associated with the task. The interaction time might be reduced by providing the participants a draggable line instead of a slider. In future work, it would also be interesting to capture the number of readjustments of the estimated lines to compare it with the response time and the accuracy results.

In both estimation and validation, the critical points are well above the 95% confidence interval, meaning that people are not able to perceive regression models with an acceptable level of accuracy by statistical standards. Therefore, individuals cannot rely on their visual validation ability. Given the relatively low self-reported expertise of the participants, the results might be different for people familiar with regression concepts, and an additional experiment solely with domain experts would be interesting for future work.

Our study design allows us to compare our results for the perception of linear trends with those of Braun et al. for the perception of the average value [30]. In general, both visual validation and estimation are less accurate for linear trends than for average values. This is probably due to the increased difficulty of the task, as the average value only examines the characteristic of one variable, while the linear trend examines the relationship

between two variables. Nevertheless, the relation between validation and estimation remains similar, but only with respect to ODR regression. The reason for this is most likely that the vertical regression is both statistically and visually decisive when calculating and perceiving the average value.

The results of the experiment showed that the participants in our study were biased towards trend lines that were "too steep". In the estimation task, they consistently drew lines with slopes that are too high, while in the validation task they accepted significantly more lines that were too steep than lines that were too flat with respect to the true trend line. Possible reasons, such as confirmation bias (i.e., given a trend line, participants may identify points that support the trend line as a valid one) [63] or the overestimation of large values and underestimation of small values [124] could not be verified by our study data and therefore require further investigation. However, the bias is reduced if the results are considered in relation to the ODR instead of the OLS regression. This indicates that, without any context or other assistance, people intuitively estimate and validate an orthogonal regression (ODR) instead of a vertical one (OLS) in scatterplots. That means, people naturally perceive errors in both variables even though they are not present in the data. As OLS is the more commonly used model for regression, we conducted a second experiment that uses additional visual augmentations to the scatterplots to improve the validation quality for OLS regression (see next section).

## 5. Experiment 2: Visual Validation with Visual Designs

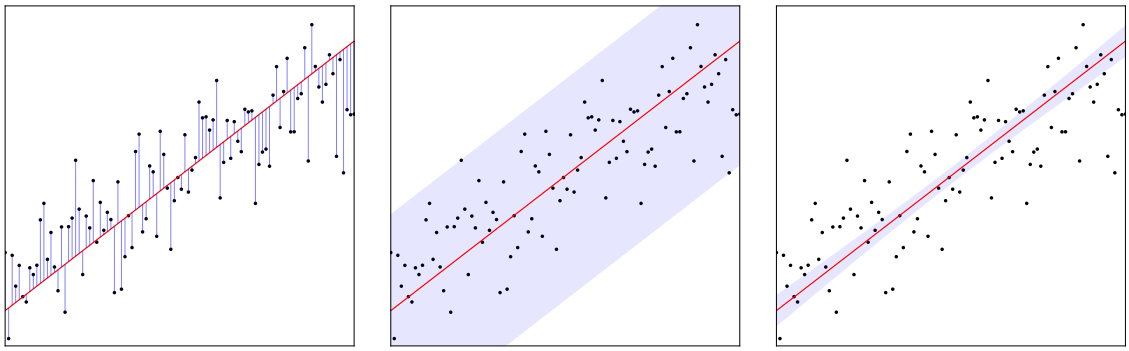
The results from experiment 1 showed that visual validation is less accurate than visual estimation. Moreover, people more likely assess orthogonal distances between data points and trend line (i.e. ODR) instead of vertical distances (i.e., OLS). However, since OLS is the more commonly-used regression model, in this section, we investigate whether the addition of visualization designs may help improve the participants' accuracy when performing visual validation with OLS regression.

Experiment 2 evaluated three common visual augmentations for regression visualization (error lines, bounding boxes, and confidence intervals), as shown in Figure 12. As described later in Subsection 5.1, we expect error lines to have the most influence on the results. Thus, our hypotheses are:

- **H1**: *Visual designs improve the validation of OLS.* It means that people accept more statistically valid and reject more invalid trend lines with respect to OLS regression.
- **H2**: *The visual design using error lines removes bias towards higher slopes.* This means, there is no perceptual difference between positive and negative slope deviations with respect to OLS regression anymore.
- **H3**: *Error lines reduce the time and difficulty of the task with respect to the other designs including the unaugmented chart.*

### 5.1 Visual Designs

Visual designs for regression validation mean additional graphical elements that are shown to the user in addition to the trend line in the visualization. We consider common visual designs for showing regression results found in literature (e.g., [138, 140, 198]) and the strategies used by participants in experiment 1 (see Subsubsection 4.2.5).



(A) OLS error lines. (B) Bounding box. (C) 95% confidence interval.

FIGURE 12: Visual designs evaluated for visual validation in experiment 2.

**Error Lines** The first visual design employs error lines [198]. Error lines show the the vertical distance of each data point from the displayed line, emphasizing the error minimized by OLS regression (see Figure 12a). This potentially relieves the user of one step of the regression calculation, which should reduce participants' response times. Moreover, by visually guiding towards OLS regression, the explicit error lines should make people perceive less the ODR regression. Therefore, it should also reduce the bias toward lines that are too steep.

**Bounding Box** The concept of bounding box covering the data was mentioned by several participants in the self-reported strategies and is inspired by Yang et al. [198]. Our implementation constructs the box by moving two lines parallel to the shown line outward until they reach the outermost points. The surface is then colored with an alpha value of 0.1 for a lower opacity (see Figure 12b). The resulting box highlights the slope of the line shown. This enlarged area, which includes all data points, is intended to help participants compare the slope of the line shown with the true trend of the data as a whole.

**Confidence Interval** Visualizing the confidence interval of a regression model is common in several different areas and applications, such as pandemic infection projections or weather forecasts [139, 156, 201]. It shows the uncertainty in the underlying model [138, 140]. All models within the confidence interval should be considered valid in a statistical sense. For our stimuli, we rotate the 95% confidence interval of the true OLS regression model in the same way as the shown line (see Figure 12c). The goal is to enhance the perception of slope deviation of a line by showing all statistically valid models that accordingly have an even greater deviation.

### 5.2 Experimental Setting and Participants

We conducted a second user study on Limesurvey [115] (with participants from Prolific [145]) using the identical data and study procedure as in experiment 1. The between-subject groups were now defined by the three designs instead of the two tasks in the first study. All participants performed the visual validation task.

After filtering for attention checks, a total of 108 participants were included in the analysis (32 for *error lines*, 38 for *bounding box*, and 38 for *confidence interval*). The demographic characteristics of the participants were similar to experiment 1: 87% were between 20 and 40 years old with 47% females and 52% males. The education levels were

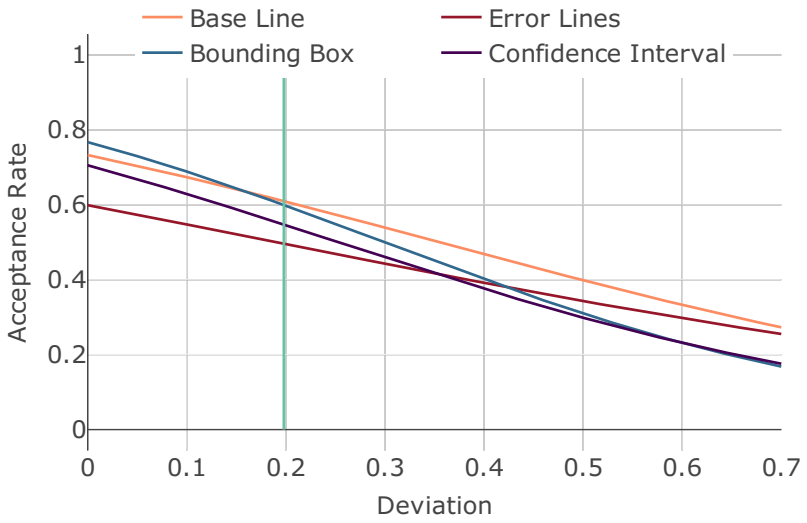



FIGURE 13: Comparison of validation accuracy (absolute deviation) with respect to OLS regression for the visual designs. Green line: Statistical 95% CI.

| **p-value** | *base* | *error* | *box* | *conf* |
|---|---|---|---|---|
| *base* | - | | | |
| *error* | < 0.01 | - | | |
| *box* | < 0.01 | < 0.01 | - | |
| *conf* | < 0.01 | < 0.01 | < 0.01 | - |

TABLE 1: p-values of the pairwise KS-test for the analysis of the acceptance rates of OLS regression.

also similar: 43% of the participants had a bachelor's degree and 82% responded with an expertise between 1 and 3 on the Likert scale.

Participants were compensated with £3.54. The average study completion time was 17 minutes.

### 5.3 Results

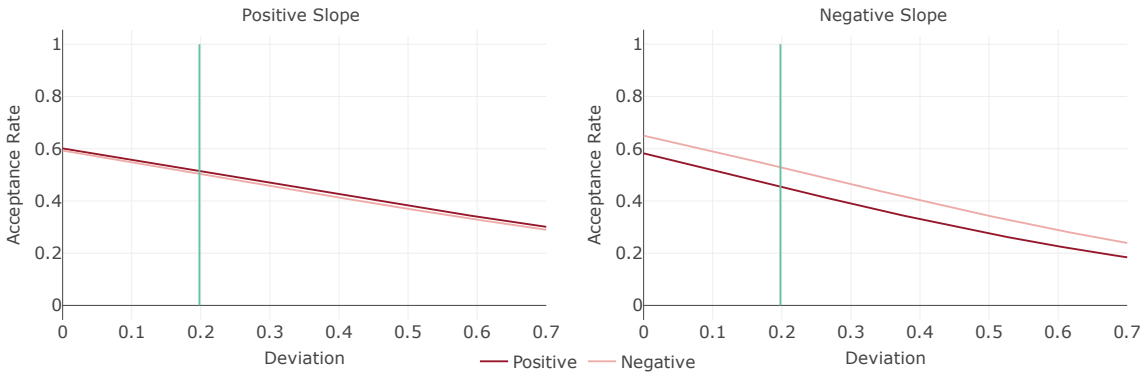
We can directly compare the results for the visual designs with the *base line* (i.e., lines without additional augmentations) from experiment 1.

**Accuracy** With respect to OLS regression (see Figure 13), the pairwise KS-test indicated significant differences in the acceptance rates between all designs (Table 1).

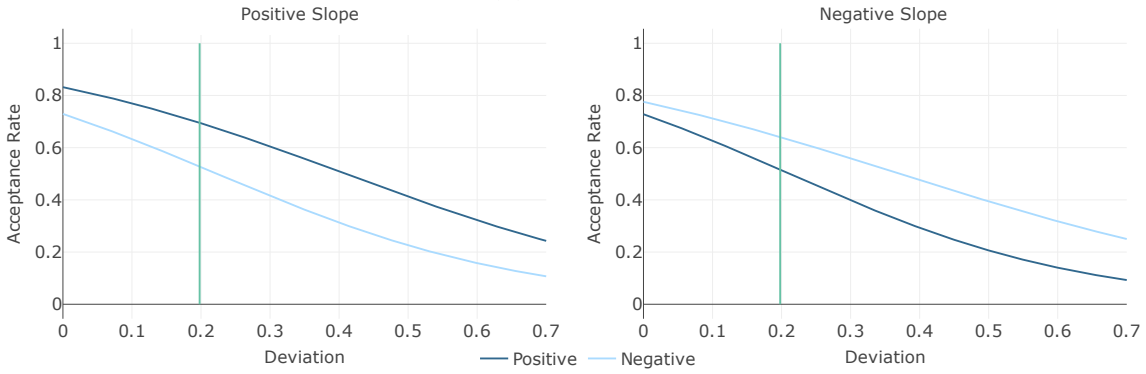
For the *bounding box* and the *confidence interval*, a similar number of valid lines (i.e., trend lines within the CI) were accepted compared to the *base line* condition, while a slightly higher amount of invalid lines were rejected. For the *error lines*, fewer lines were accepted that were within the statistical 95% confidence interval. The true trend lines were accepted only 60% of the time.

The overall decrease in the acceptance rates with visual designs have lowered the critical point between acceptance and rejection of a shown line for all three designs ($crit_{error} = 0.190$, $crit_{box} = 0.251$, $crit_{conf} = 0.298$).

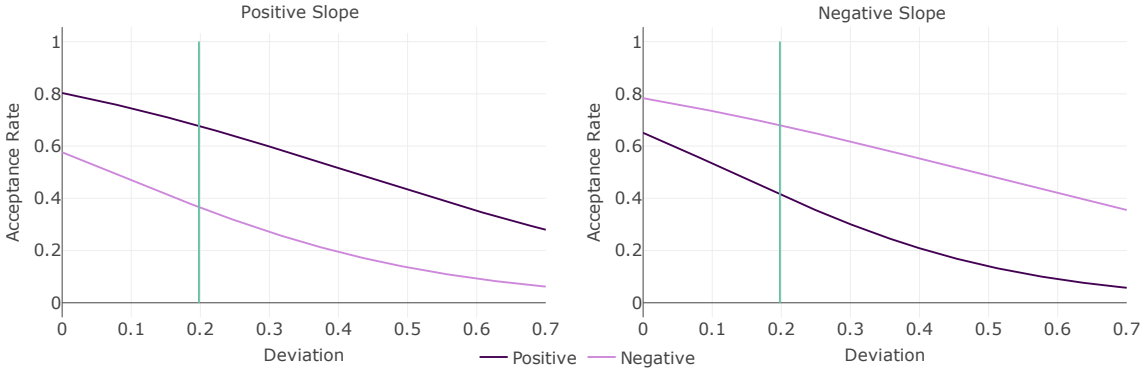
In sum, people rejected slightly more invalid models with visual designs, but did not accept more valid models. Therefore, our hypothesis **H1** cannot be rejected.




(A) Error Lines.

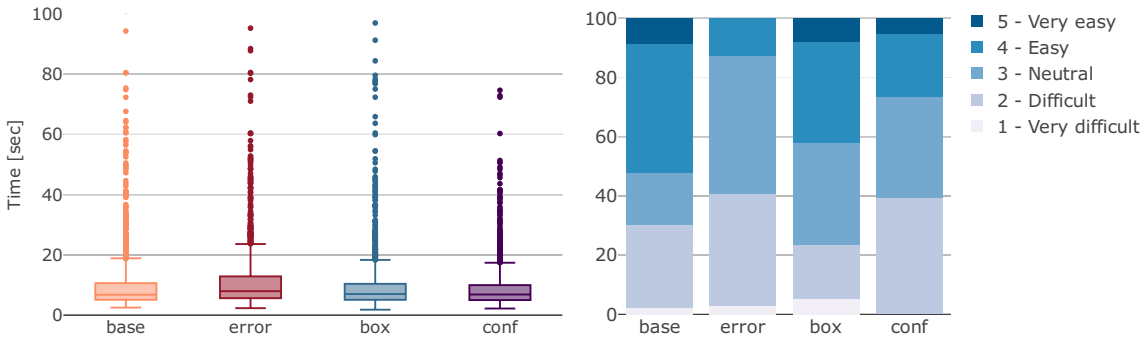



(B) Bounding Box.




(C) Confidence Interval.

FIGURE 14: Comparison of the validation acceptance rates for the different visual designs for positive and negative deviations and for positive and negative trends with respect to OLS regression. Green line: statistical 95% CI.




(A) Response times. (B) Self-reported task difficulty.

FIGURE 15: Response times and self-reported task difficulty for the different visual designs in experiment 2.

**Bias** Figure 14 shows the differentiated acceptance rates for the designs. For the *error lines*, the deviation difference between the acceptance' critical points of positive and negative slope deviation decreased ($\Delta^{pos}_{error} = 0.020$ for positive trends and $\Delta^{neg}_{error} = 0.117$ for negative trends). While the acceptance rates for negative trends still differ (KS: $p < 0.01$, $D = 0.182$), they are not significantly different for positive trends anymore (KS: $p > 0.71$, $D = 0.044$). This result suggests that the perception of OLS regression was improved for the participants with *error lines*. It partially supports hypothesis **H2**.

Similar to the *base line*, the *confidence interval* biased the perception of positive and negative slope deviation ($\Delta^{pos}_{conf} = 0.348$, $\Delta^{neg}_{conf} = 0.353$). The *bounding box* showed a slight improvement in the perception of lines that were too steep ($\Delta^{pos}_{box} = 0.186$, $\Delta^{neg}_{box} = 0.159$).

**Response Time and Difficulty** The response times (measured in seconds) and participants' self-reported task difficulty on a 5-point Likert scale are shown in Figure 15.

The Kruskal-Wallis test indicated a significant difference in the response times ($p < 0.01$, $\chi^2 = 112.42$). A pairwise Wilcoxon test of the visual designs (Table 2a) showed that the response times with the *error lines* were significantly higher than for all other designs ($\mu_{error} = 12.64sec$). Among the other designs, the *confidence interval* had the fastest response times ($\mu_{conf} = 9.21sec$), followed by the *bounding box* ($\mu_{box} = 9.96sec$) and the *base line* ($\mu_{base} = 10.85sec$).

A significant difference in the designs was also found in the perceived difficulty (Chi-squared: $p < 0.05$, $\chi^2 = 21.19$). The task was perceived to be the easiest with the *base line* ($\mu_{base} = 3.28$). The only significance could be found when comparing the *base line* to the *error lines* (Table 2b) where the *error lines* was perceived as the most difficult ($\mu_{error} = 2.69$). In between the *base line* and the *error lines* were the *bounding box* ($\mu_{box} = 3.21$) and the *confidence interval* ($\mu_{conf} = 2.92$).

Summarizing the results, hypothesis **H3** is not supported.

### 5.4 Discussion

The results of experiment 2 show that people consistently reject trend lines more often when they are presented with visual designs regardless of whether the trend lines are valid or invalid. There could be several different reasons for this: 1) The visual designs help to identify incorrect regression models. 2) People are generally more skeptical when shown model results with visual designs because they are not used to or do not understand this type of presentation. 3) People perceive the type of model to not fit to the underlying data. In our study, we intentionally did not provide an explanation for the designs to allow for a purely perceptual study. It is therefore possible that a prior

| **p-value** | *base* | *error* | *box* | *conf* |
|---|---|---|---|---|
| *base* | - | | | |
| *error* | < 0.01 | - | | |
| *box* | 1.00 | < 0.01 | - | |
| *conf* | 0.92 | < 0.01 | 0.58 | - |

(A) Response times.

| **p-value** | *base* | *error* | *box* | *conf* |
|---|---|---|---|---|
| *base* | - | | | |
| *error* | < 0.01 | - | | |
| *box* | 0.38 | 0.06 | - | |
| *conf* | 0.11 | 0.36 | 0.18 | - |

(B) Self-reported difficulties.

TABLE 2: p-values of the pairwise Wilcoxon-test for the response time analysis and pairwise chi-squared test for analysis of the self-reported difficulties.

explanation of the visual designs would improve participants' understanding and thus the results. Similarly, the results may be influenced by the low statistical expertise of our participants.

With respect to OLS regression, the acceptance thresholds improved with all designs compared with the *base line*. This is due to slightly higher rejection rates of invalid trend lines. The validation accuracy of valid lines did not improve with any design.

As with the *base line*, there is no bias in trend direction with the visual designs. Although confidence intervals are commonly used to represent statistical uncertainty, the bias in slope deviation is greatest with the addition of the confidence interval in the visualization. The error lines, on the other hand, provided an unbiased validation in terms of "too steep" and "too flat" trend lines.

The addition of *error lines* in the visualization should in theory reduce the cognitive effort of visual validation, because the cognitive calculation of errors is not needed. However, participants took longer to complete the validation task and found it more difficult than with the *base line*, according to the results of our study. This suggests that people either did not fully understand the concept of error lines without explanation, that the processing of additional information is cognitively demanding, or that it forces people to intensify their thinking about the shown line and to correct their bias.

The results of experiment 2 showed that the addition of commonly used visual designs for visualizing regression in a scatterplot does not significantly improve people's ability to validate models visually. Therefore, we are unable to provide design guidelines. As visual estimation remains more accurate, this suggests that guiding people to cognitive estimation as a first step of visual validation might improve accuracy.

## 6. Limitations and Future Work

We investigated the visual estimation and validation of linear trend lines in scatterplots. The findings and limitations in our experiments may suggest new research questions and future directions.

**Model complexity:** We found that individuals' ability to visually validate a linear model is lower than their ability to validate a constant model (i.e., averages [30]). This raises the question whether the ability to visually validate is dependent on the complexity of a model. In order to answer this question, however, it would be necessary to define model complexity in relation to visual perception.

**Data characteristics - Outliers:** Our study analyzed data with normal distribution as assumed by OLS regression models. However, real world data sets may have special characteristics that impact the regression. For example, Correll and Heer [52] studied the influence of noise and outlier in the data. The addition of outliers would shift the validation question of a correct or incorrect model to a question of including or excluding the outliers in the regression. This question about the correctness of the model itself could be extended to a study on the validation of model types. In this, people would have to be decided whether the type of model fits the data or not. In this paper, we examined the fit of the parameters of a fixed model.

**Visual designs for model validation:** We tested four designs (unmodified base line and three visual design augmentations) that are commonly used in visualizing regression models. The addition of the three visual designs in experiment 2 failed to improve both the accuracy and bias of visual validation. It would be interesting to see whether the results improve in a separate study where explanation and context to the data and the visual designs are made available to the participant. Additional think-aloud sessions could further provide insights into people's visual validation process and the use of visual designs. Altogether, the development of novel visual designs for model validation that improves both accuracy and mitigates bias is a future challenge.

## 7. Conclusion

In summary, our research examines the effectiveness of visual validation in assessing linear regression models shown in scatterplots. We conducted two empirical experiments to gain insight into individuals' abilities to visually validate linear trends and the impact of common visualization designs on validation quality.

The first experiment showed that participants were more accurate at visually estimating linear trends in scatterplots than at visually validating them. In addition, our results revealed a bias in slope deviation (i.e., toward slopes that are "too steep"), but no bias in trend direction. Additional analysis provides evidence that people naturally assess orthogonal regression (ODR) rather than the most commonly used vertical regression (OLS). This indicates that people assume errors in both variables rather than in just the y-coordinate.

The second experiment aimed to evaluate whether incorporating common visualization designs such as error lines, bounding boxes, and confidence intervals could improve visual validation. Despite the reduction in validation bias observed with error lines, none of the tested designs yielded the desired improvements in accuracy.

Our results emphasize the limitations of relying solely on visual model validation for linear regression models in scatterplots. Further research is needed to investigate the underlying cognitive processes involved in visual validation tasks in order to find appropriate visual solutions for supporting visual model validation.

## Acknowledgments

The authors would like to thank all study participants. This paper is a result of Dagstuhl Seminar 22331 "Visualization and Decision Making Design Under Uncertainty". This work has been partially supported by the BMBF WarmWorld Project, the SANE Project, and the Risk-Principe Project. This work has also been funded in part by NSF Awards 2007436, 1452977, and 2118201.

## 3.4 Visual Validation of Linear Trends With Outliers

In the previous work, the model complexity was increased from the average value to linear trends. For this project on visual model validation, the data complexity is increased this time. The paper investigates the extent to which individuals take outliers into account when validating linear trends in scatterplots. To this end, the influence of different outlier factors is examined and the validation results again are compared with the results of visual estimation.

The paper is in preparation for submission to CHI conference on Human Factors in Computing Systems:

**D. Braun**, D. Eberle, R. Chang, M. Gleicher, and T. von Landesberger. Deciding Through Noise: Visual Validation of Linear Trends in Scatterplots Amid Outliers. 2026.

The supplementary material of the paper, including the study preregistration, data, results, and documentations, as well as the Python code for the data and stimuli generation, is publicly available at OSF.

I am the shared first author of this work together with Dominik Eberle. In this role, I am mainly responsible for the design, implementation, data collection and analysis, as well as the writing of the paper. The specific contributions of myself and my co-authors to this work are outlined below according to the Contributor Roles Taxonomy (CRediT):

**D. Braun**: Conceptualization, Methodology, Data Curation, Formal Analysis, Investigation, Software, Visualization, Writing – Original Draft. *D. Eberle*: Methodology, Data Curation, Formal Analysis, Software, Visualization. *R. Chang, M. Gleicher, T. von Landesberger*: Supervision, Conceptualization, Methodology, Writing – review & editing.

# Deciding Through Noise: Visual Validation of Linear Trends in Scatterplots Amid Outliers

Daniel Braun[*,1], Dominik Eberle[*,1], Remco Chang[2], Michael Gleicher[3], Tatiana von Landesberger[1]

[*]*Shared first authors*
[1]*University of Cologne*
[2]*Tufts University*
[3]*University of Wisconsin-Madison*

**Abstract:**

We investigate the influence of outliers on the visual validation of linear trends. When some data points do not follow the core trend, different statistical models can be applied depending on the treatment of these data points. When a viewer is asked to validate a linear regression, they must make an assumption about how to weight the outliers. Through two quantitative human-subject studies, we investigate which models participants adopt when visually validating regression models and how various factors influence their decisions. The first study compares visual model validation to the related task of visual model estimation. The results show that when validating, viewers are less consistent about how to account for outliers than when estimating, and that validation is less robust to data factors. A second study explores validation behavior in more depth. We find that individual viewers consistently accept models that either include or exclude outliers, regardless of whether the presence of outliers is mentioned or not. These decisions are influenced solely by outlier congruence with the trend direction. Our findings improve the understanding of visual model validation of linear trends in scatterplots with outliers, offering practical implications for machine learning applications and the design of visual analytics systems.

*In preparation for submission to CHI conference, 2026*

## 1. Introduction

Visual validation of statistical models is a core part of exploratory data analysis [49,96] and visual analytic systems [12,42,165,192]. Given the inherent complexity of many models (e.g., regressions), visual inspection is often employed to assess their accuracy and reliability [40, 41]. This process is especially critical when model outputs have substantial implications in sensitive areas such as pandemic forecasting [55] and meteorological prediction [102]. Traditional statistical metrics may fail to capture the nuanced characteristics of the underlying data and model, as data sets with divergent properties can yield similar numerical statistics [21, 122, 181]. Then, visual validation comes to its strength. It reveals subtle patterns of the data and models.

Outliers – data points that deviate markedly from a core trend – pose significant challenges in statistical analysis and modeling [13, 120]. They may arise from measurement errors, data entry mistakes, or noise, or they can be correct but just unusual [203]. A variety of models and methods has been developed to deal with such data points. In the context of ordinary least squares (OLS) regression, two of the most straightforward models involve either [52]: (A) including all observations – thereby assuming that the outliers reflect legitimate variation in the data – or (B) classifying the data that does not fit as outliers and ignoring them (see Figure 1). As both models are statistically plausible, it is up to humans to choose which of the two to use. This can be done visually by looking at the data and its fit to the calculated model.

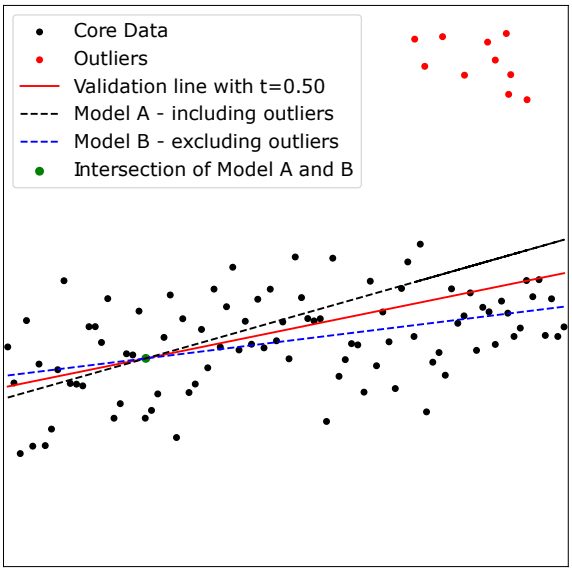



FIGURE 1: Example scatterplot containing outliers with the two statistically valid OLS regression models: The model including the outliers (*model A*) and the model only considering the core data (*model B*). The red line shows an example validation line for a value of $t = 0.5$ for the slope $s = t \cdot a + (1 - t) \cdot b$ with $a$ and $b$ being the slopes of model A and B.

For visualization designers, it is important to know which of the valid models viewers naturally assess. This affects which models to display in a visualization or whether there is a need to give the viewer visual cues. These insights are critical because effective data visualization does more than present information -— it shapes perception and cognition, guiding decision-making in scientific and applied contexts [57, 58, 127, 178].

The research on the visual communication of statistical models has been focused on two tasks: *visual model estimation* [46, 52, 80, 85, 95, 106, 126, 132, 136, 151, 152, 176, 177, 196, 198, 199] – individuals' ability to visually fit a model to data – and *visual model validation* [27, 30] – individuals' ability to assess the fit of a given model to the underlying data. Braun et al. [27] found significant differences in individuals' performance in visual validation and estimation of linear regression models in scatterplots. These studies were performed on data without outliers. It is unclear how people validate models in the presence of outliers, which frequently occur in real scenarios and data and are not trivial to analyze.

In two human-subject studies, we investigate the behavior of viewers during visual validation of linear model results when outliers are present in the data. As outliers can occur in various ways (e.g., in different quantities or at the beginning or end of a trend), we examine possible factors that influence the data (see Figure 2). In line with the literature [45, 52], these data factors include the outlier quantity, trend direction, horizontal positioning of the outliers, and outlier congruence to trend direction. Congruent outliers lie on the side of the core trend that supports this trend (i.e., incongruent outliers lie on the contradictory side).

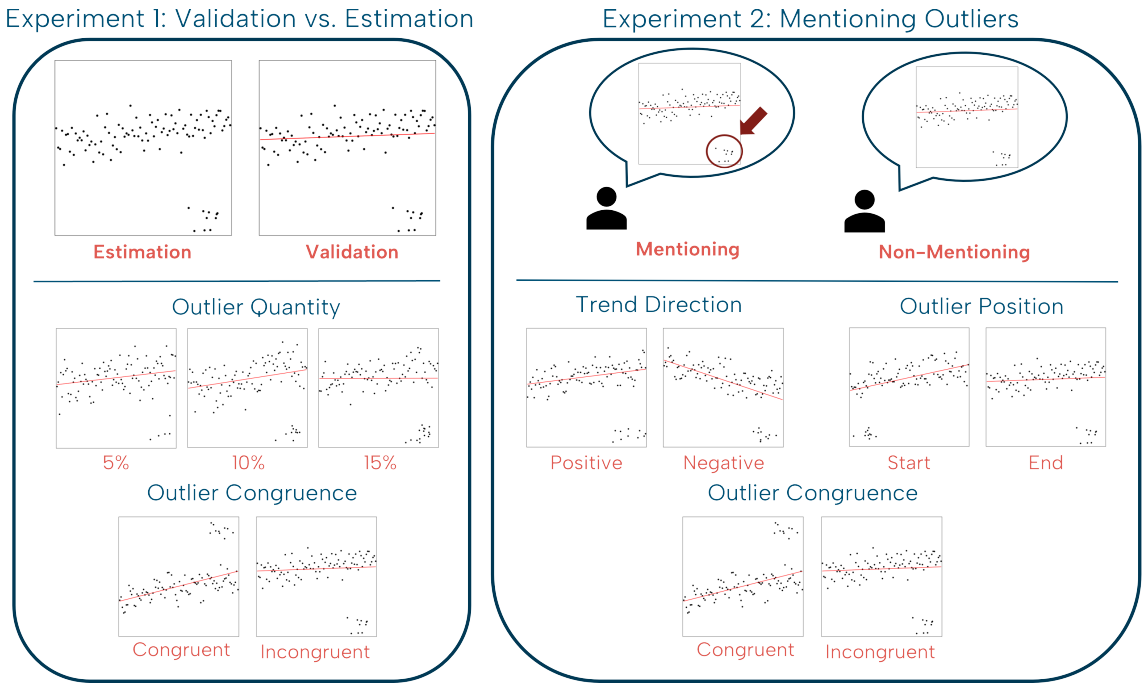



FIGURE 2: Experimental factors tested in our human-subject studies on the visual validation of linear trends containing outliers. The first experiment compares visual model validation and estimation with varying outlier quantity and congruence of their position with the trend direction. In experiment 2, we examine the difference in visual validation between mentioning and non-mentioning the outliers to participants including the factors: horizontal positioning of the outliers, vertical congruence to the trend direction, and the trend direction itself.

The first experiment compares visual validation with visual estimation of linear trends with outliers. In a between-subjects setting with volunteer participants, they saw a scatterplot containing outliers and had to either accept or reject given trend lines (validation) or adjust the slope of a line to fit the trend of the data (estimation). We found significant differences in the validation and estimation behavior of the participants. Moreover, the results show significant variations in the tendency to weight the outliers during validation for the tested factors outlier quantity and outlier congruence.

In the second experiment, we examined whether explicitly mentioning the existence of outliers causes a difference in the consideration of outliers in visual validation. A difference in validation behavior between mentioning and non-mentioning of outliers would have implications for visualization designs, e.g., the phrasing of chart titles or annotations. We found no difference in the results between mentioning and non-mentioning, either overall or for the individual data factors, indicating that people are not biased by task phrasing. Outlier congruence is also the only factor that changes the validation tendency, whether or not outliers are included. Comparing the individual results of each participant indicates that visual validation with outliers is individual and a more cognitive than perceptual task.

Our work deepens the understanding of how individuals interpret anomalous data points in visual model validation tasks. The insights gained can inform the design of visual analytic systems that aim to bridge the gap between complex statistical outputs and intuitive human understanding. Our findings provide a foundation for the development and future research of the visualization and annotation of model results that supports

human perception and decision-making. Further research is needed on the cognitive factors involved in the visual validation process when outliers are present, such as context and trust.

The paper is organized as follows: Section 2 reviews current research on visual estimation and visual validation. Section 3 details the overall data, study, and analysis design. The results of the two experiments and their discussions are presented in Section 4 and Section 5. Section 6 addresses the limitations of the current work and outlines potential avenues for future research.

## 2. Related Work

**Visual Model Estimation** The visual estimation of statistical models has been widely researched. Besides works on the estimation of average values [80, 95, 126, 199] and correlation [85, 106, 152, 176, 177, 181, 196, 198], many papers exist on the estimation of linear trends.

Correll and Heer [52] conducted a series of crowdsourced experiments to assess the perceptual processes of linear trend estimation using "regression by eye" across multiple bivariate data visualizations. They found that the accuracy of the participants' estimated trends with respect to OLS regression could be significantly influenced by both visual features and data features, such as outliers. Ciccione and Dehaene [46] and Liu et al. [116] conducted similar experiments with the results that people consistently overestimate trends and that trend estimation is sensitive to the absolute slope value.

In their estimation with outliers experiment, Correll and Heer [52] found that people estimate trend lines significantly closer to the OLS model excluding outliers than the model including them. Their results did not indicate a significant effect with respect to the outlier position. Increasing the outlier quantity did not influence trend estimations. The studies of Filipowicz et al. [73] support that people give outliers less weight than the other data when visually estimating. Notably, Ciccione et al. [45] came to a different conclusion: In their outlier experiments, participants did not exclude outliers in their estimations. Even when they were mentioned and specifically asked to reject in the trend estimation, the outliers still had some impact, although the participants were able to accurately detect outliers. In their studies, the distance of the outliers to the main data as well as the outlier quantity influenced viewers' mental regressions. In the experiments of Oral and Boduroglu [137], participants weighted outliers equally to the non-outliers, and even overweighted them when those outliers fitted a certain context given to the participants. In our work, we specifically avoid context in order not to influence the perception.

All of these works were important references for our study, the hypotheses and the design of the stimuli, including the influencing factors that we tested. In addition, we can compare the results of our first experiment with the literature.

**Visual Model Validation** The perceptual and cognitive characteristics of visual validation have not been studied as extensively as estimation.

Braun et al. [30] were the first who investigated the perceptual differences between visual validation and visual estimation of average values in scatterplots. They found a significant difference between the processes with participants being more accurate in estimation than in validation. In a follow-up work they increased the model complexity and examined the visual validation of linear trends in scatterplots [27]. Although the overall performance in both visual validation and estimation decreased, estimation remained

more accurate. Moreover, the authors found a "too steep" bias in visual validation, i.e., participants more likely accepted trend lines that were steeper than the true trend. Standard approaches for showing the model results (e.g., confidence intervals and error lines) did not increase participants' validation performance.

These studies motivated our work on the validation of linear trends in scatterplots containing outliers. The addition of outliers makes the validation task and its investigation more complex, as there is no longer one simple, statistically correct model. In addition, the relevance for applications is higher, as outliers are regularly included in real-world data.

Other works consider visual model validation mostly as an interactive part of visual analytics or machine learning systems [22,40,42,43,130]. Although our focus is on perception and cognition instead of interactive systems, our work deepens the understanding of visual model validation, aiding the future development of visual analytics and machine learning systems.

An important part of model validation is the viewer's trust in the model and its visualization [41]. In line with existing literature on trust in visualization [56,66,123,142,154], we minimized the influence of trust by using minimalistic visualizations and providing participants with only the necessary information to understand the tasks.

## 3. Experimental Design

In two experiments, we investigated humans' perceptual behavior when visually validating linear trends in scatterplots containing outliers.

Both studies contain the same *visual validation task*: Participants viewed scatterplots (containing outliers) with a pre-drawn trend line (see Figure 2) and were asked to either accept or reject the line as a fit for the data's trend.

In the **first experiment**, we compare visual validation with visual estimation. The task formulation for validation was: "How does the slope of the drawn line relate to the linear trend of the scatterplot?" and participants had the response options "too steep", "too flat", and "about the same". To estimate a trend line, participants were asked to adjust the slope of a line by moving a slider to fit the trend of the data in the scatterplots (task formulation: "Please indicate the slope of the linear trend of the scatterplot."). The rotation point of the line to be estimated was fixed at the intersection of the models including and excluding the outliers, so that participants had the possibility to estimate either model. This technique was already used in previous works on visual estimation [27,52].

The **second experiment** examines the difference in people's validation behavior when the presence of outliers is explicitly mentioned. Thus, the study included two different task formulations, for which participants had to answer with "accept" or "reject":

- *Non-Mentioning*: "Does the shown trend line fit the data?"
- *Mentioning*: "Does the shown trend line fit the data containing **outliers**?"

### 3.1 Study Procedure

The two experiments were conducted online and created with Limesurvey [115]. Although they were designed to address different research questions, they shared a similar study structure.

Both studies used a *between-subject design* with two groups each. They were defined by the validation and estimation tasks in the first experiment and by mentioning and not mentioning the outliers in the second experiment. The between-subject study design prevented learning effects and reduced the number of trials per participant [38]. The same data were used for the between-subject groups to ensure consistency and comparability.

The same *study procedure* was used in both experiments: After asking demographic questions about the participants, we gave them a short training period to familiarize themselves with the study interface. We did not provide training feedback to avoid bias in the participants' responses. The actual study trials were randomized to minimize trial ordering effects. Each page in the study interface contained one trial (i.e., one plot) and the response times per trial were recorded. After the completions of the main tasks, we asked the participants to rate the difficulty of the tasks on a 5-point Likert scale [186] and for their strategy in considering the outliers (in free text form).

### 3.2 Data Generation and Stimuli Design

Our data generation combines and adapts the approaches of Correll and Heer [52], Braun et al. [27], and Liu et al. [116]. Each scatterplot contained 100 data points in the range of $[0,1] \times [0,1]$ and had a size of $700 \times 700$ pixels, which is why we recommended a screen size of 13" or larger. In the following, a distinction must be made between the core data and the outliers.

**Core Data** The sets of core data points were generated using standard ordinary least squares (OLS) regression models: $y = ax + b$. To ensure comparability of the trends, we bounded the slope of the regressions to a range of $a \in \{x \in \mathbb{R} \mid 0.1 \leq |x| \leq 0.2\}$. This was also necessary in order to create a sufficient distance between the outliers and the core data. For the same reason, we have shifted the y-coordinate of the central point $x = 0.5$ to either = 0.4 or 0.6 (depending on whether the outliers were above or below the core trend). The y-intercept $b$ was subsequently calculated by the fixed center point and the randomly chosen slope value $a$. The resulting data points were uniformly distributed along the x-axis to avoid overplotting. Their y-coordinates were subsequently permuted using Gaussian distribution with a fixed standard deviation (0.1 in the first experiment, 0.075 in the second experiment) for a similar noise level [151]. With 100 data points in total, the number of core data points is $n_{\text{core}} = 100 - n_{\text{outliers}}$.

After the first experiment, we adjusted the core data generation for the second experiment by adding additional constraints: We reduced the standard deviation (see above) and all core data points were allowed to have a maximum distance of twice the standard deviation from the underlying core data trend line. This was done because it is important that only the outliers we provide are perceived as outliers if they are explicitly mentioned in the question, as we want to analyze how participants deal with them.

**Outliers** As there is no consensus definition for visual outliers, we tested several approaches from previous works [45, 53, 165]. Since it allows us to best control outliers for our study conditions, we decided to adapt Liu et al.'s [116] outlier clustering method and Correll and Heer's [52] spatial separation (see Figure 3). For an appropriate outlier distinguishability, we placed the outlier cluster center $(x_c, y_c)$ in the x-range $[0, 0.3]$ (beginning) or $[0.7, 1]$ (end) and y-range $[0, 0.1]$ (bottom) or $[0.9, 1]$ (top) dependent of the outlier position (i.e., top or bottom 10% and left or right 30% of the plot). The respective outliers were then computed by:

$$x_i = (x_i - x_c)\delta + x_c \tag{1}$$

$$y_i = (y_i - y_c)\delta + y_c \tag{2}$$

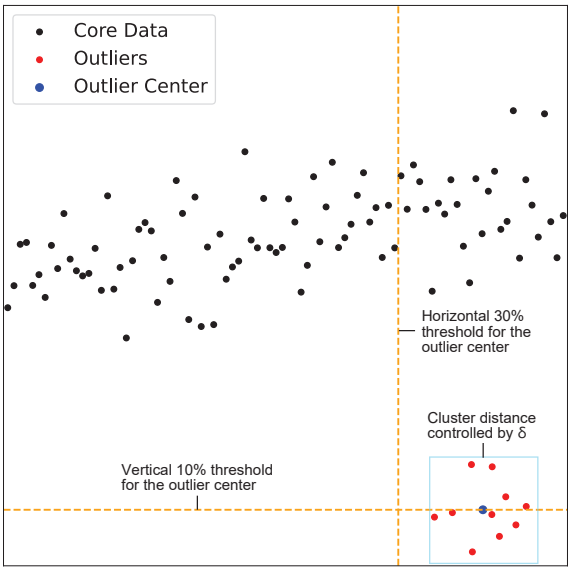



FIGURE 3: The outlier generation we used for our data creation adapted and combined from Correll and Heer [52] and Liu et al. [116].

The parameter $\delta$ controls the distance between the outliers in the cluster. For equal conditions across trials, we set $\delta = 0.5$ after visual inspection. We added random noise to the x- and y-coordinates sampled from a Gaussian distribution with half the standard deviation of the core trend.

**Slope (t-Values)** The slope value $a$ is the variable we investigate in our studies. Therefore, we deviate the slopes of the shown lines to assess the participants' behavior. We fixed the intersection point between models A (including the outliers) and B (excluding the outliers) for each line, so that the slope is the only independent variable. The slope of the shown line can than be described as a linear combination of the two OLS regressions:

$$\text{shown slope} = t \cdot \text{slope}_A + (1 - t) \cdot \text{slope}_B \tag{3}$$

We can control the slope of the validation line by alternating between different values of $t$. This also allows us to assess participants' tendency towards either model, as a value $t = 1$ results in model A and $t = 0$ in model B. Figure 1 shows an example of a validation line with $t = 0.5$. In both studies, we test the t-values:

$$t \in \{-0.25, 0, 0.17, 0.33, 0.5, 0.67, 0.83, 1, 1.25\}$$

**Data Factors** In the two experiments, we test several additional factors defined by specific data and outlier characteristics in addition to the between-subjects groups. The goal is to investigate which of these factors influence people's validation behavior. Figure 2 gives an overview of the considered factors in both studies:

- **Outlier quantity**: The proportion of outlier points in the total number of points: {5%, 10%, 15%} *(exp 1)*
- **Outlier congruence**: The vertical positioning of the outlier in relation to the trend direction: {congruent, incongruent} *(exp 1 & 2)*

- **Outlier position**: The horizontal positioning of the outlier points: {beginning, end} of trend *(exp 2)*
- **Trend direction**: The orientation of the core trend: {positive, negative} *(exp 2)*

As an example for the congruence of the outliers, for a positive trend, outliers at the end that are above the core data are congruent and those that are below the core data are incongruent (see Figure 2). The opposite is true for outliers at the beginning.

### 3.3 Analysis Procedure

Similar to the analysis of Braun et al. [27,30], we transformed participants' validation responses into binary results: *1* for *accepting* the shown line, *0* for *rejecting* it. Subsequently, we calculated the acceptance rate (mean validation response of all participants) for each t-value and applied polynomial regression of 4th-order on these rates. The 4th-order regression includes the possibility of having two maximum values in case participants accept both the model including ($t = 1$) and excluding ($t = 0$) the outliers and reject other t-values. To analyze the individual data factors, we aggregated the results for the other factors. To compare the results of validation and estimation, the visually estimated t-values were calculated using the same definition (Equation 3) and 4th-order polynomial regression was applied on the number of estimations per t-value.

In our statistical analysis, we conducted all tests with a significance level of $\alpha = 0.05$. Initially, we applied the Shapiro-Wilk test to both the responses and response times to assess normality; however, none of the data met the normality assumption. Consequently, we opted for non-parametric methods. We used the Kolmogorov-Smirnov (KS) test to compare the validation and estimation results (i.e., comparing the polynomial regressions of the validation acceptance rates and the estimation values). We applied the Wilcoxon test to evaluate differences in the means of response times, and the chi-squared test for the Likert-scale responses.

## 4. Experiment 1: Visual Validation versus Visual Estimation

The first experiment was designed to investigate whether visual model validation and estimation behave differently for linear trends in scatterplots containing outliers. We evaluated whether different data characteristics influence the results to answer the following research questions:

- **RQ1.1**: Do people consider outliers when visually validating linear trends in scatterplots or not (if we don't tell them about outlier existence and do not tell them what to do)?
- **RQ1.2**: How does the validation behavior differ from the behavior in the visual estimation of linear trends lines of data containing outliers?
- **RQ1.3**: Does the visual validation and behavior change for specific data characteristics (i.e., outlier quantity and outlier congruence)?

The hypotheses developed from these research questions are based on our own assessments during data generation and previous research on model perception [27,30,45,52]:

- **H1.1**: *People behave differently for visual validation and visual estimation of linear trends containing outliers.*
  The studies by Braun et al. [27,30] showed significant differences in the results of the two tasks. We expect this result to hold with outliers.

- **H1.2**: *With an increased outlier quantity, participants are more likely to accept trend lines including the outliers.*
  We expect participants to assign more importance to outliers in their visual validation with higher outlier ratio, similar to earlier research findings on visual estimation of trend lines with outliers [45, 52].

- **H1.3**: *The congruence of the outliers with the trend direction influences the participants' validation behavior.*
  Based on the results of Braun et al. [27] that lines steeper than the true trend were more likely to be accepted, we expect this effect to be amplified by the addition of outliers.

- **H1.4, H1.5**: *The {completion time, self-reported difficulty} is lower for visual validation than for visual estimation of linear trends containing outliers.*
  Whereas validation task is a matter of acceptance, the estimation task requires participants to mentally fit a line to the data. Therefore, we expect the results from previous research [27] to be confirmed.

### 4.1 Experimental Setting and Participants

In order to limit the number of trials in the first study, only the outlier quantity and congruence were included as data factors in addition to the validation and estimation distinction. The direction of the trend was set to be positive and outliers were only shown at the end of the trend.

The participants of the first experiment were recruited through advertisements in lectures and via the platform SurveyCircle [179] and did the study voluntarily. The study contained 54 trials and the average completion time was 15 minutes. From 82 participants in total, six participants were filtered out for failing attention checks. The remaining 76 participants were equally divided between the validation and estimation groups. The gender distribution was almost balanced (F: 55%, M: 45%) and the majority were between 20 and 30 years old (88%). Their educational levels ranged from high school diplomas to masters degrees. The self-reported expertise in statistical model estimation was roughly normally distributed on a 5-point Likert scale ($\mu = 2.7$) and there was no statistical correlation between the expertise and the two groups ($p = 0.22$). Thus, the validation and estimation results are comparable.

### 4.2 Results

We compare the distribution of acceptance rates per t-value with the estimated t-values (see Subsection 3.3) and analyze changes in the validation distributions based on the data factors. Moreover, we compare response times and task difficulty and summarize participants' self-reported strategies in solving the tasks.

**Difference Between Visual Validation and Visual Estimation** In contrast to previous work on visual model validation [27, 30], we do not measure the accuracy of validation and estimation in terms of a correct regression model. Since there are two statistically valid regressions in the presence of outliers (including and excluding the outliers), we instead examine which model people tend to validate and estimate.

Figure 4 compares the overall distributions of validation acceptance rates and estimated lines summarized for the data factors (outlier quantity and congruence). The t-value indicates the (relative) slope deviation of the shown lines, with a t-value of 0 meaning excluding the outliers and a t-value of 1 meaning including them in the regression

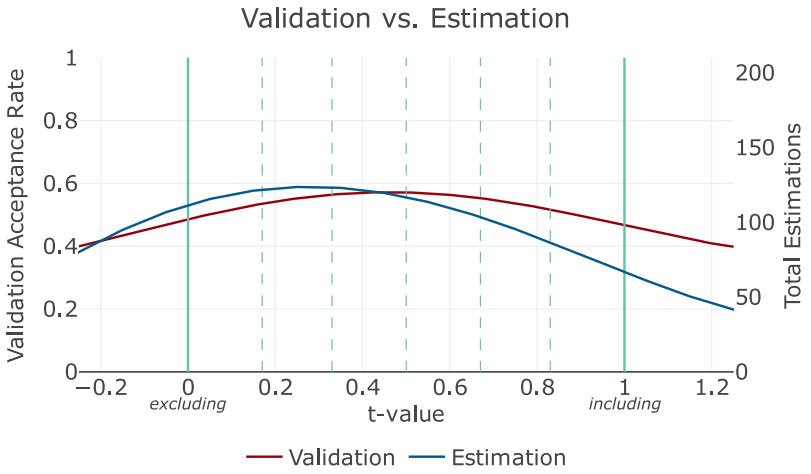



FIGURE 4: Comparison of validation acceptance rates with estimated lines in experiment 1. As shown in this figure, participants considered outliers differently when performing visual validation and estimation.

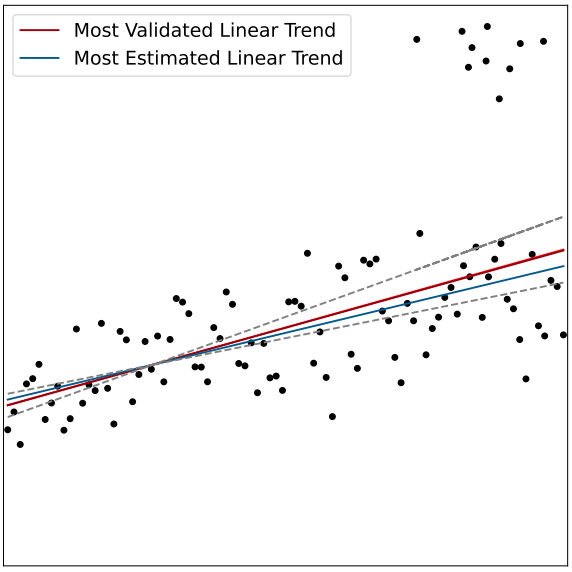



FIGURE 5: Summarizing example of the results for visual validation and estimation in experiment 1. The figure shows the most accepted and estimated trend lines. The dashed lines display the OLS regressions including and excluding the outliers. Notably, outliers are more likely to be considered in visual validation than in estimation.

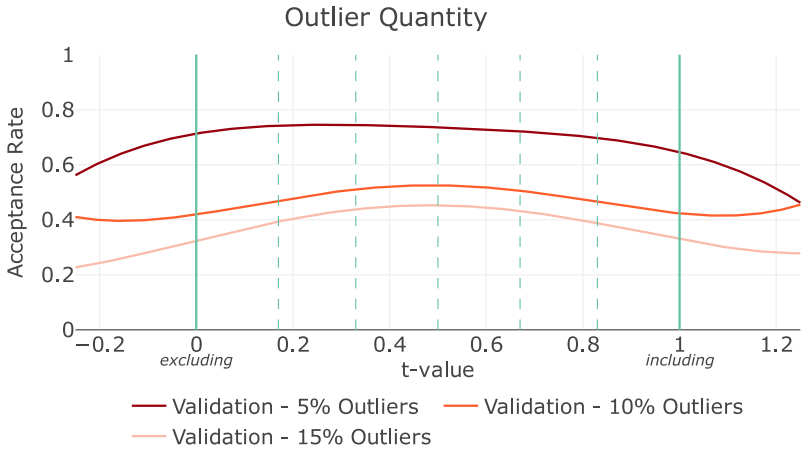



FIGURE 6: Comparison of validation acceptance rates for different outlier quantity in experiment 1. As shown in this figure, participants considered outliers differently across outlier quantities when visually validating.

model. If the participants would only accept or estimate the two actual OLS models, the polynomial regressions of the results would have to have two maxima at these t-values.

As shown, the distribution of the visually estimated trend lines peaks at t-value = 0.25 with 124 estimations. This skewed distribution close to the t-value 0 indicates that participants tended to exclude outliers in their estimations. In contrast, the distribution of acceptance rates appears to be more balanced, peaking at t-value = 0.46 with an acceptance rate of 57%. According to this, participants most frequently accepted lines that lie exactly between the two actual OLS models. An example of the most validated and estimated trend lines is visualized in Figure 5. It is also noticeable that the acceptance rates of the visual validation condition toward t=1 do not decrease as much as the number for the visual estimation condition and remain above 40%.

The Kolmogorov-Smirnov test confirms the statistical difference in the validation and estimation distributions with a p-value $< 0.01$. Therefore, hypothesis **H1.1 is accepted**.

**Influence of Outlier Quantity** The comparison of the validation acceptance rate distributions for different outlier quantities (5%, 10%, and 15% outliers) is displayed in Figure 6. Notably, the overall acceptance rates drop with an increased number of outliers. The polynomial regressions between the three conditions do not intersect for any quantity. This fact is confirmed by the maximum values of the distributions: With 5% outliers, the acceptance rates are the highest with a maximum of 75% for t-value = 0.27. It is followed by a maximum of 53% for t-value = 0.49 with 10% outliers. With 15% outliers, the participants accepted the least amount of lines (45% for t-value = 0.48).

The distributions for 10% and 15% outliers are very similar. For these outlier quantities, people are more likely to consider the outliers, and are most likely to accept lines that lie between Model A and B. For 5% outliers, the maximum is closer to t-value 0, indicating that the participants were more likely to exclude the outliers. The pairwise KS tests show that the difference in the 5% outlier acceptance rate distribution from the other two is statistically significant (Table 1). The distributions for 10% and 15% outliers do not differ significantly. Thus, **H1.2 is partially accepted**.

The estimation distributions are similar for all three outlier quantities in the way that people tend to exclude outliers ($\max_{5\%} = 0.25$, $\max_{10\%} = 0.25$, $\max_{15\%} = 0.35$). They do not significantly differ.

| **p-value** | *5%* | *10%* | *15%* |
|---|---|---|---|
| *5%* | - | | |
| *10%* | <0.01 | - | |
| *15%* | <0.01 | 0.29 | - |

TABLE 1: p-values of the pairwise Kolmogorov-Smirnov tests for the analysis of the acceptance rates for different outlier quantities in experiment 1.

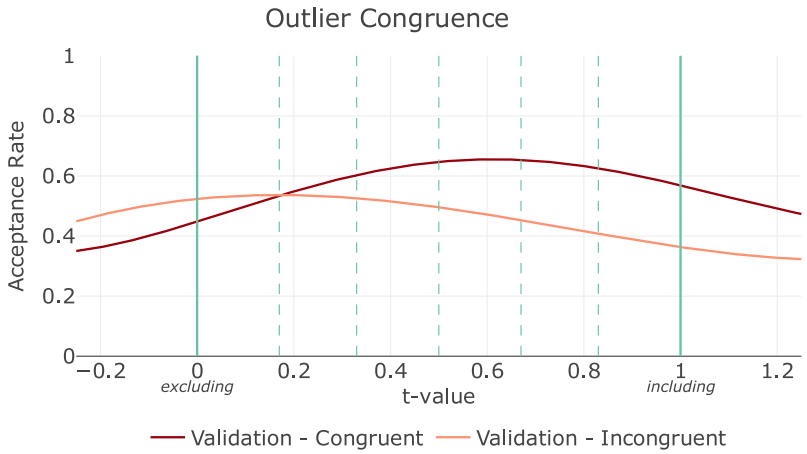



FIGURE 7: Comparison of validation acceptance rates for different outlier congruence with trend direction in experiment 1. As shown in this figure, participants considered outliers differently based on outlier congruence when performing visual validation.

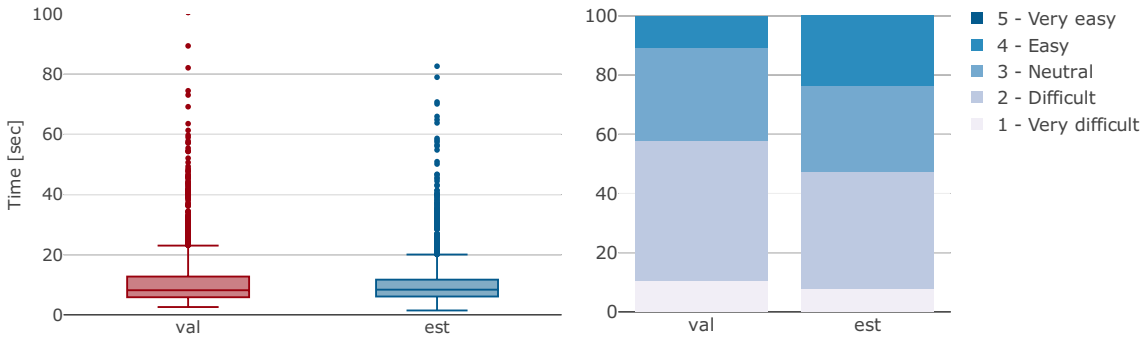
**Influence of Outlier Congruence** With the slope direction and horizontal position of the outliers fixed to positive trends and outliers at the end, the outlier congruence is defined by their vertical position: outliers above the core trend are congruent, outliers below incongruent.

Comparing the validation acceptance rates for the two characteristics in Figure 7 shows a clear difference in their distributions. Trend lines with congruent outliers have slightly higher overall acceptance rates with the maximum of 66% at t-value = 0.61. In contrast, the highest acceptance rate for incongruent outliers is 54% at t-value = 0.17. This means that participants tend to include congruent outliers in the validation process, while they do the opposite for incongruent outliers. The statistical significance of this difference is confirmed by the Kolmogorov-Smirnov test ($p$ = 0.01), which is why **H1.3 is accepted**.

Again, this factor does not have a significant effect on the participants' visual estimation, as they tended to ignore the outliers in both cases ($\max_{\text{congruent}}$ = 0.35, $\max_{\text{incongruent}}$ = 0.25).

**Response Time and Difficulty** The response times per trial and self-reported task difficulty are displayed in Figure 8. Although the mean response times are slightly different ($\mu_{val}$ = 12.01, $\mu_{est}$ = 10.45), the Wilcoxon test shows no significant difference between the two tasks ($p$ = 0.39). Thus, hypothesis **H1.4 is not accepted**.

Moreover, **H1.5 is not accepted**, since the chi-squared test showed no significant dependence between task and difficulty level ($p$ = 0.50).




(A) Response times. (B) Self-reported task difficulty.

FIGURE 8: Response times and self-reported difficulty in experiment 1.

**Self-Reported Strategies** Regarding the overall strategies, two participants of each group reported to have tried to balance out the number of data points above and below the proposed or estimated linear trend line. With regard to the strategy for considering outliers, more participants in both groups stated that they had included the outliers (7 for validation, 8 for estimation) than ignoring them (3 for validation, 5 for estimation).

In the case of outliers' positioning, four participants in the validation group and five participants in the estimation group reported giving congruent outliers, which appeared to be closer, more weight. For the outlier quantity, the difference in participant feedback was higher between the two groups, with eight participants for validation and only four participants for estimation giving outliers more significance when the quantity of outliers is higher.

### 4.3 Discussion

*Validation vs. Estimation:* The results of experiment 1 show that viewers behave significantly different when visually validating and estimating linear trends in scatterplots containing outliers, providing a clear answer to **RQ1.2**. Our estimation results confirm the findings of Correll and Heer [52] of users' estimates being closer to the outlier-excluding model B. For visual validation, the participants showed no overall tendency towards either including or excluding the outliers (**RQ1.1**). In fact, they accepted the lines with slopes exactly between the two actual OLS models most often. People were inconsistent in how to consider the outliers when shown various trend lines as they were not given any information about the context of the data. This motivates our second experiment whether it makes a difference if the participants are explicitly told that outliers are present in the data.

The results of Braun et al. [27] showed that participants have a higher accuracy when visually estimating linear trends, than when validating them. Thus, people may intended to include outliers, but were inaccurate in doing so. In the feedback regarding their overall strategies, most participants reported to have included the outliers in their evaluation of the proposed trend lines in some way. On the other hand, the feedback of the estimation group indicates that most users tried to include outliers as well, which confirms the findings of Ciccione et al. [45] regarding observers not spontaneously rejecting outliers in their visual estimations when not explicitly told to. However, the actual results of the estimation group suggest otherwise, as participants tended to exclude outliers in their estimations. This indicates that there is a confirmation bias [63, 172] toward including outliers in a statistical regression with visual validation that is not present when visual trend estimation is left to viewers.

*Number of Outliers:* Our results show that outlier rates of 10% or higher do not significantly influence the validation tendency of the participants toward including or excluding them. With 5% outliers, however, people tend to exclude them in their validations. This is supported by the reports from participants that they gave more weight to outliers with an increased quantity. Though, the t-values measure the relative slope deviation for the shown trend lines (see Figure 1). With an increased number of outliers, the deviation in slopes between model A and B increases as well. Therefore, participants accepted a higher absolute slope deviation, but it did not increase relatively with the outlier quantity.

It is also noticeable that the acceptance rates for visual validation decreased for all t-values as the number of outliers increased. This could be reasoned by a higher uncertainty about the "correct" linear trend with an increased outlier ratio due to a higher absolute slope deviation between the two OLS models, leading to a higher rejection rate overall. Consistent with Correll and Heer [52], the outlier quantity did not have an effect on the visual estimations.

*Congruency:* In their feedback on their strategies, many participants reported to have given congruent outliers, which were positioned above the main data trend, a higher weight. They appeared to be less far from the main trend, or seemed to fit the overall trend more accurately than bottom outliers, as the core data always followed a positive trend. These reports on the relevance of the outliers' position fit our results, as the participants considered congruent outliers more often in their validation of linear trends. Notably, these findings enhance the "too steep" bias found by Braun et al. [27]. For bottom outliers, however, the opposite is the case, as most accepted linear trends were closer to model B. Thus, observers are less likely to be influenced by outliers that deviate too much from the core trend.

*Summary:* Results for the data factors and answering **RQ1.3** show that participants' validation behavior was influenced by outlier quantity and congruence. For visual estimation, however, the distribution of estimated trend lines did not change regardless of the outlier properties. This means that visual model estimation is robust against outlier characteristics and people consistently exclude outliers in their estimations. Our estimation results confirm the findings of Correll and Heer [52].

## 5. Experiment 2: Mentioning Outliers and Additional Validation Factors

In the second experiment, we test the effect of explicitly mentioning the presence of outliers to participants as well as additional data and outlier characteristics on visual validation. In experiment 1, we did not give participants any context about the data or the outliers. In this study, we make participants aware of the existence of outliers, but not explicitly tell them to include or exclude them. A difference in validation behavior between mentioning and non-mentioning of outliers would have implications for visualization designs, e.g., the phrasing of chart titles or annotations. The two between-subjects groups consist of mentioning and non-mentioning the outliers in the task description to answer the research questions:

- **RQ2.1**: Does the behavior for visual validation of linear trends with outliers differ if we raise awareness of the outliers (without telling the participants what to do with them)?
- **RQ2.2**: Does the validation behavior change for either case (mentioning and non-mentioning) with specific data characteristics (outlier position, trend direction, outlier congruence)?

After a pilot study with 20 participants (10 per group), the experimental plan including hypotheses, sampling, and analysis was preregistered on OSF[1]. Our hypotheses for the study are based on the results of experiment 1, the pilot study, and previous research [27, 45, 52]:

- **H2.1**: *Participants behave differently for visual validation of linear trends with outliers when they are explicitly informed of the presence of outliers.*
  We expect the participants to make a decision whether to exclude or include the outliers for the trend validation when they are explicitly mentioned. Accordingly, the results would differ from those of the non-mentioning group.

- **H2.2**: *In both cases, the tendency of the participants towards either including or excluding outliers in visual validation is not influenced by the outlier position.*
  Correll and Heer [52] found no significant difference in visual estimation with varying horizontal outlier positioning. We expect this result to hold for visual validation.

- **H2.3**: *In both cases, the tendency of participants towards either including or excluding outliers in visual validation is not influenced by the trend direction.*
  Visual validation behavior of linear trends without outliers was not influenced by the trend direction [27]. There is no expectation that the addition of outliers will change this.

- **H2.4**: *In both cases, participants behave differently for visual validation of linear trends with outliers, depending on whether the outliers are positioned congruently or incongruently to the trend direction.*
  The first experiment showed a significant influence of outlier congruence to trend direction. We expect this effect to persist even if the direction of the trend changes.

- **H2.5, H2.6**: *The {completion time, self-reported difficulty} is lower for visual validation when the presence of outliers is explicitly mentioned.*
  If the outliers are mentioned, we assume that the participants will decide in advance how they will deal with the outliers. Without mentioning the outliers, they might rethink them with each trial.

### 5.1 Experimental Setting and Participants

The second experiment contained the factors: outlier position, outlier congruence, and trend direction. As there was no difference in the distribution of acceptance rates with 10% outliers or more, we fixed the number of outliers at this ratio. To determine whether participants were able to make sense of the information about the outliers, we also added a question asking whether they knew the meaning of the outliers (after the experiment, so as not to prime them).

The participants for the second study were recruited on Prolific [145]. The average completion time for the 72 trials was 17 minutes. The participants were compensated with 4.30€ for completing the task. After filtering for participants who failed the attention checks (7 participants), the responses of 80 participants were included in the analysis (40 per group). Similarly to the first study, the gender distribution was balanced (F: 49%, M: 50%, D: 1%) and most participants had a university degree (bachelor: 46%, master: 29%). Other educational levels were high school graduate (16%), vocational training (8%) and

[1] https://osf.io/j64pr/?view_only=e627f6d538d1444b99ee0a4ad06f071b

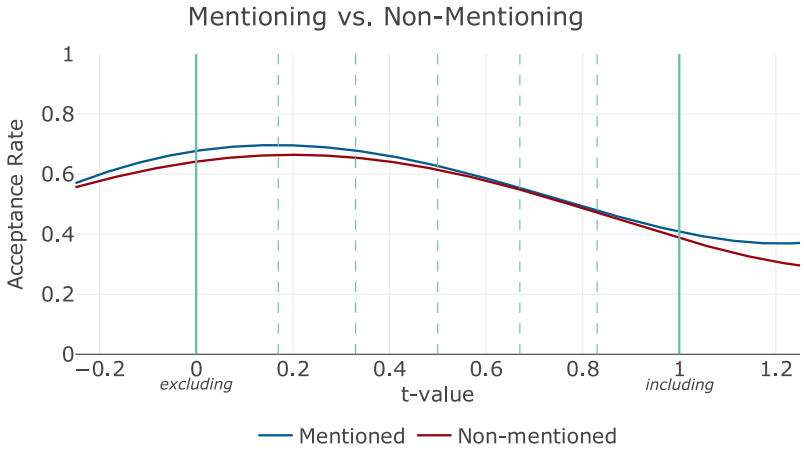



FIGURE 9: Comparison of validation acceptance rates for mentioning and non-mentioning the outliers in experiment 2. As shown in this figure, participants in the mention and the non-mention groups considered outliers similarly when performing visual validation.

nursery school to 8th grade (1%). The participants' age ranged from 20 to 60 (20-30: 55%, 31-40: 31%, 41-50: 11%, 51-60: 2%) and 86% reported an statistical regression expertise of 3 or lower ($\mu = 2.3$). Again, the chi-squared test indicated no correlation between the expertise of the participants and the two between-subject groups ($p = 0.39$), so that the mentioning and non-mentioning results are comparable. In total, 85% of the participants knew the meaning of outliers (90% in the mentioning group, 80% in the non-mentioning group).

### 5.2 Results

**Difference Between Mentioning and Non-Mentioning Outliers** Figure 9 compares the acceptance rates for mentioning and non-mentioning the outliers summarized for all participants and all data factors. Both distributions look very similar and their peaks are very close to each other. For the mentioning group, the maximal acceptance rate of 70% is reached at $t = 0.16$. In the case of non-mentioning, the maximum is at an acceptance rate of 66% at $t = 0.20$. In both groups, the acceptance rates remain above 60% around t-value 0 and fall to 40% around t-value 1. These right skewed distributions indicate that the participants tended to accept lines more likely that had a slope more similar to the OLS regression model excluding the outliers. Moreover, it makes no difference whether people are made aware of the presence of outliers or not. The Kolmogorov-Smirnov test confirms that the two distributions differ only insignificantly ($p = 0.97$). Therefore, hypothesis **H2.1 is not accepted**.

**Influence of Outlier Position** The comparison of the acceptance rates differentiated by the outlier positions in Figure 10 shows no major differences between the two groups. Only with outliers at the end of the trend there are minimal changes: for non-mentioning, the overall acceptance rate drops slightly, and for mentioning, the peak shifts a bit toward $t = 0$. Otherwise, the acceptance rates follow the summarized distributions. This means that participants more likely exclude outliers in their validation, regardless of whether they are at the beginning or end of the trend (respectively left or right of the plot). This is also not changed by mentioning the outlier. The Kolmogorov-Smirnov test ($p_{\text{men}} = 0.99$, $p_{\text{non}} = 0.85$) confirms that **H2.2 is not rejected**.

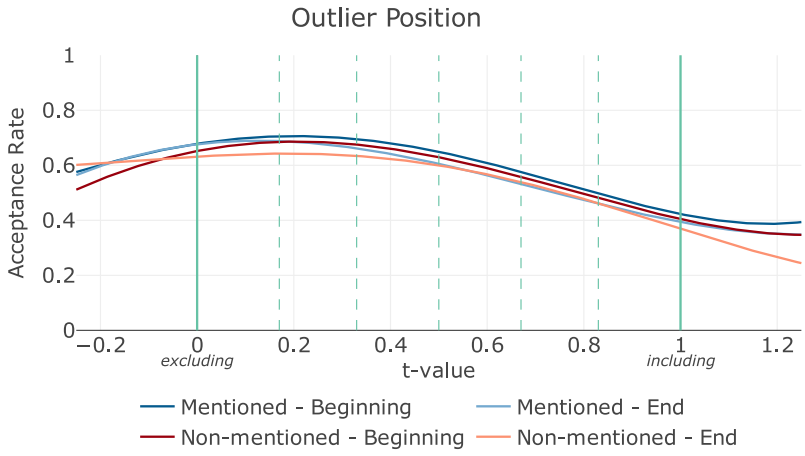



FIGURE 10: Comparison of validation acceptance rates for different outlier positions in experiment 2. As shown in this figure, participants considered outliers similarly across outlier positions.

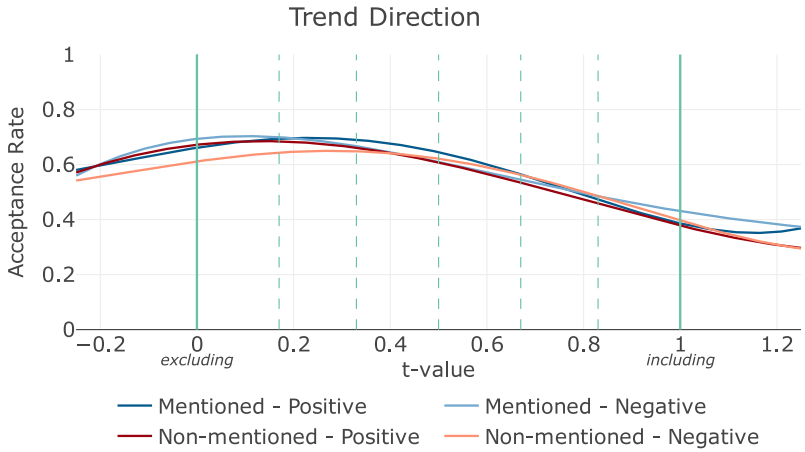



FIGURE 11: Comparison of validation acceptance rates for different trend directions in experiment 2. As shown in this figure, participants considered outliers similarly across trend directions when performing visual validation.

**Influence of Trend Direction** The acceptance rates per trend direction show a similar picture as the outlier position (see Figure 11). The distributions remain similar, with a higher acceptance rate at $t = 0$ and a drop at $t = 1$. This means that the tendency of participants to exclude outliers persists. The maxima of the curves shift slightly more than in the outlier position (mentioning: $t_{\text{positive}} = 0.23$, $t_{\text{negative}} = 0.10$; non-mentioning: $t_{\text{positive}} = 0.13$, $t_{\text{negative}} = 0.28$). Nevertheless, the Kolmogorov-Smirnov test shows no significant differences for the groups ($p_{\text{men}} = 1$, $p_{\text{non}} = 0.80$), so that the trend direction does not influence peoples' validation behavior and **H2.3 is not rejected**.

**Influence of Outlier Congruence** Congruent outliers in positive trends are outliers are below the core trend at the beginning and above the core trend at the end. For negative trends this is reversed. For this factor, the distributions of the acceptance rates for the two characteristics vary, although they are similar for the two between-subjects groups (see Figure 12). For incongruent outliers, the participants accepted most often the OLS regression excluding the outliers ($t_{\text{men}} = 0.06$, $t_{\text{non}} = 0.09$). The acceptance rates

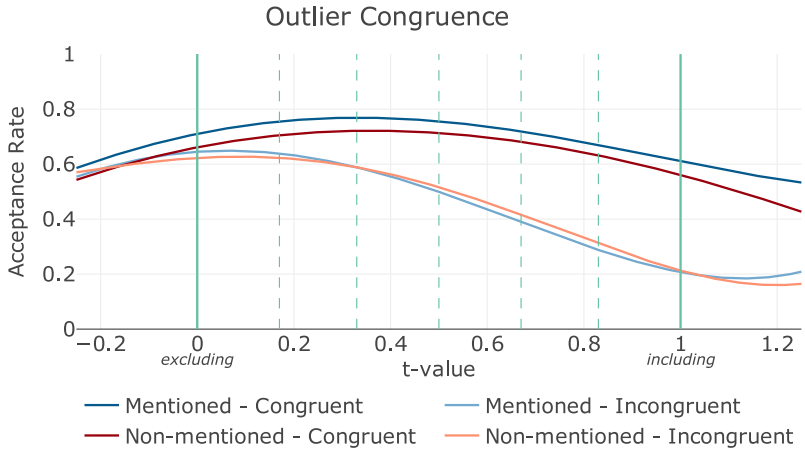
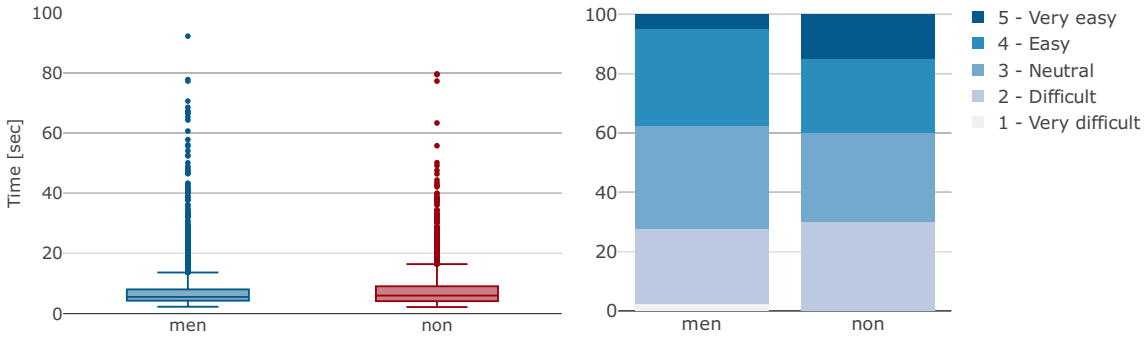



FIGURE 12: Comparison of validation acceptance rates for different outlier congruence with trend direction in experiment 2. As shown in this figure, participants considered outliers differently based on outlier congruence when performing visual validation.



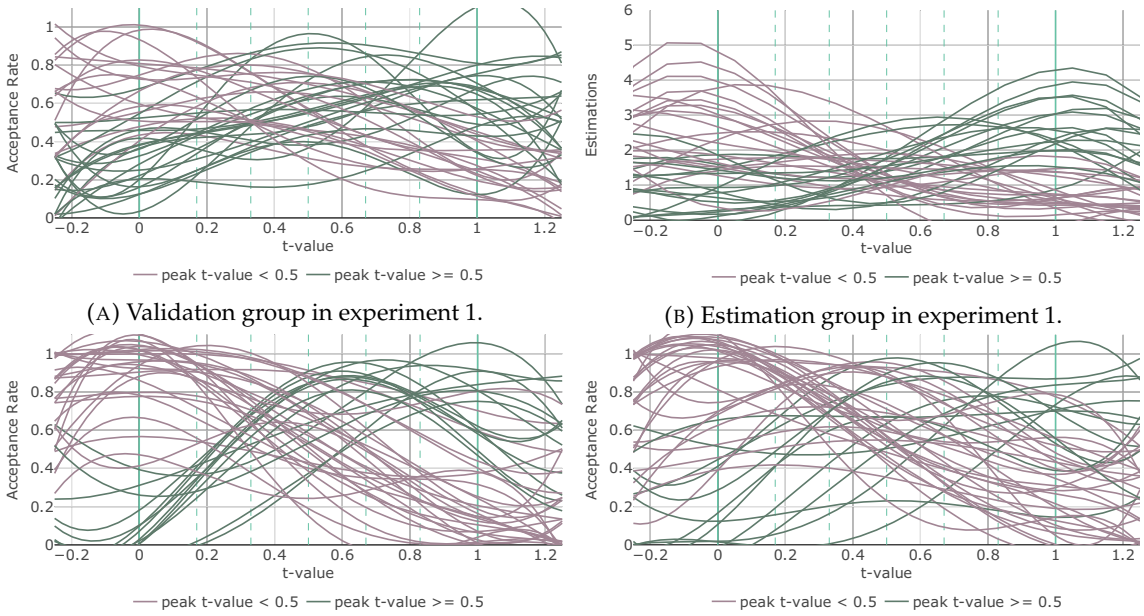

(A) Response times.

(B) Self-reported task difficulty.

FIGURE 13: Response times and self-reported difficulty in experiment 2.

drop sharply to an acceptance rate of only 20% as more outliers are considered. For congruent outliers, however, all slopes were consistently accepted at least 60% of the time. Here, the peak t-value shifts more toward 1 ($t_{\text{men}} = 0.33$, $t_{\text{non}} = 0.36$), indicating that the participants gave more weight to the outliers in their validations. According to the Kolmogorov-Smirnov test, the distributions for both groups are significantly different ($p_{\text{men, non}} < 0.01$) and **H2.4 is accepted**.

**Response Time and Difficulty** Figure 13 shows the response times and the self-reported task difficulties by group. The median response time is lower for the mentioning group ($\tilde{x} = 5.54$) than the non-mentioning group ($\tilde{x} = 6.00$). The Wilcoxon test confirms ($p = 0.03$) that **H2.5 is accepted**. The self-reported task difficulty results show the opposite effect, with the mean for non-mentioning ($\mu = 3.25$) being slightly higher than for mentioning ($\mu = 3.13$). Here, the chi-squared test shows no significant difference between the two groups ($p = 0.44$), so that **H2.6 is not accepted**.




(A) Validation group in experiment 1.

(B) Estimation group in experiment 1.

(C) Non-mentioning group in experiment 2.

(D) Mentioning group in experiment 2.

FIGURE 14: Participants' individual estimations and acceptance rate distributions for experiment 1 and 2 (each line represents one participant). Notice that for the estimation in experiment 1 (b) and the validation in experiment 2 (c,d), there is a clear individual consistency to either considering or not considering the outlier.

### 5.3 Additional Analyses

Participants in the mentioning and non-mentioning groups seem to behave very similarly, which is a finding contrary to the results of experiment 1 and our initial hypotheses. To get more insights into the differences between the two groups, we conducted further exploratory investigations beyond the preregistered analyses.

A plausible explanation for the discrepancy is the different participant pools of the two studies. The first study mainly involved volunteers and students, the second study was conducted by paid study takers.

To get an insight into the difference between participants, we look at the individual results of each participant. The individual distributions indicate that individuals consistently either include or exclude outliers across trials. This is evident in both the non-mentioning validation in experiment 2 (see Figure 14c) and the estimation in experiment 1 (see Figure 14b), as as there are two clearly separable clusters of curves. Notably, participants who include the outliers estimate the OLS regression more accurately (most peak t-values around 1) than they validate it (most peak t-values around 0.6).

Mentioning the outlier seems to create inconsistencies among those people in their decision to exclude the outliers. This can be seen by the fact that the distributions with peak t-values greater than 0.5 vary more than before (see Figure 14d). For the validation group in experiment 1, this division of participants is not that clear (see Figure 14a). In all cases, the visual validation when excluding the outliers was very accurate (most peak t-values around 0).

### 5.4 Discussion

*Individual Differences:* The extended analysis of experiment 2 indicates that the visual validation of linear trends with outliers depends on the individual participants. The participants seem to be divided into those who tend to include the outliers and those who exclude them. For the participants who tend to include the outliers, mentioning the presence of outliers in the study does not affect their visual validation performance. However, for the participants who tend to exclude the outliers, becoming aware of the outliers causes inconsistencies in their decisions (to include or exclude the outliers in visual validation). Overall, since more participants tend to exclude outliers, the average result skews towards a distribution that appears as if there is no difference between the two study conditions. Upon further inspection, we observe that the differences in the two study conditions do, in fact, have an effect, which provides a multi-layered answer to **RQ2.1**.

*Cognition:* The observed differences between these participants in the weighting of the outliers suggest that the task of visual validation changes from a perceptual to a cognitive one as soon as outliers are noticed. This is supported by the fact that 85% of participants stated that they knew what the term outliers means. This implies that participants decide whether or not they want to consider outliers when performing visual validation, regardless of the data conditions.

*Data Factors:* The observation that the task of visually validating trends with outliers is cognitive, is supported by the fact that most of the data factors we tested had no effect on validation behavior. Only the outlier congruence changes the acceptance rates of the shown trend lines (as in experiment 1): Outliers that follow the trend in terms of vertical positioning are given more weight in the validation than incongruent ones. The validation behavior of the participants did not differ between mentioning and non-mentioning for any factor. Therefore, the answer to **RQ2.2** is that visual validation is resilient to most data factors and is only affected by the proximity of the outliers to the core trend.

*Outlier Separability:* The importance of the separability of the outliers from the core trend is also reflected in the different results of the two experiments in the case of visual validation. In experiment 1, participants showed no tendency toward either of the two OLS regression models, while in the second experiment they formed two clusters (inclusion and exclusion of the outliers). With our additional analysis (Subsection 5.3), we have shown that the different data factors tested are not the decisive reason. This means that besides the different participant pools (with the resulting varying individual behaviors). Another possible reason is our adaption of the data and outlier generation. In the second experiment, we further increased the distance of the outliers to the core data: we reduced the standard deviation of the core trend and set a maximum allowable distance of core data points from the underlying regression. In this way, we wanted to make sure that only the outliers we intended would be perceived as outliers, so that the effect of mentioning would be reinforced. Thus, it seems that the more separate outliers are (and therefore better recognized as such), the more confident viewers are in their validation decisions about how to deal with the them. However, to prove our assumption would require an experiment specifically designed for outlier separability.

*Accuracy:* As described in Subsection 5.3, participants were very accurate at validation when excluding outliers, but not so accurate when including them. However, they were very accurate in their visual estimations in both cases. This can be explained by the results of Braun et al. [27], who showed that people are more accurate at visual estimation of linear trends than at visual validation. This imbalance seems to persist when outliers are involved.

*Summary:* Summarizing our results, we found evidence that there are two possible approaches when dealing with outliers: include/exclude (binary), or weighting (continuous). The overall results seem to suggest that participants tend to exclude outliers across all conditions: estimation, validation with mentioning, and validation without mentioning. This finding is amplified when the condition contains incongruent outliers. In the non-mentioning condition (Figure 14c) we can observe a consistent behavior across individuals. However, when outliers are explicitly mentioned (Figure 14d), their behavior becomes less consistent. Some participants stay the same (including outliers), but some start to use the "weighting" strategy instead of the binary inclusion or exclusion. In this regard, what we observe is that the "mention" condition causes an additional cognitive layer to an otherwise perceptual study (which looks at how much outliers subtly affect estimation or validation decisions when participants include or exclude outliers).

## 6. Limitations and Future Work

We examined the visual validation of linear trend lines in scatterplots that include outliers. Our experimental results, together with identified limitations, suggest several directions for future research.

**Study Design:** Our post-hoc analysis indicated an individual decision to include or exclude outliers in the visual validation. However, our study was not designed to examine individual differences, so we cannot make stronger statements about these results. An experiment explicitly designed to confirm individual differences is needed.

**Context:** In our experiments, we gave the participants as little information as possible about the data in order to focus on perception. However, previous work has shown that the context of the data can strongly influence the assessment of outliers in visual estimation [196]. We would like to find out whether the effect also occurs with visual validation.

**Trust:** Trust in the data and its source is directly linked to the data context [41, 56, 123]. Depending on the context or source of the data, the viewer may have more or less trust in it, which also influences the decision on how to deal with the outliers. The role of trust in visual validation has not yet been researched.

**Outlier Validation:** Our data and outlier generation were defined in a way so that outliers could be identified as clearly as possible. We also conducted our analyses so that only the data points we considered to be outliers were taken into account as such. However, our study design did not allow us to determine which data points the participants considered as outliers. The study by Ciccione et al. [45] shows that people are able to visually determine outliers accurately. It would be interesting to investigate how viewers would validate (i.e., to what extent they would accept or reject) given outliers (e.g., through colored data points). This could provide an understanding of the properties (e.g., distance, density, etc.) for which data points are considered outliers.

## 7. Conclusion

In summary, our work examines the behavior of individuals when visually validating linear regression models presented in scatterplots that contain outliers. We conducted two human-subject studies to gain insights into how viewers deal with outliers in their visual assessment of given linear trend lines and the impact of different data and outlier characteristics.

The first experiment showed that people behave differently with respect to outliers depending on whether they visually validate or estimate the linear trends. While participants tended to ignore the outliers in the visual estimation, outliers were given more

weight in the visual validation. Furthermore, the tested data factors had a significant impact on the validation results, while the estimation was robust to them: A higher quantity of outliers led to greater inconsistency among participants, resulting in generally lower acceptance rates for all shown trend lines. In addition, we found that outliers placed congruently with the direction of the trend were more likely to be considered in the validation than incongruent outliers.

In the second experiment, we tested the effect of explicitly informing participants of the existence of the outlier. The results showed that viewers were not biased by the mentioning of the outliers and that there was no significant difference from the results without mentioning. The factors trend direction and horizontal positioning of the outliers also did not influence participants' validation behavior. Only for the outlier congruence we saw to the same effect as in experiment 1, showing that the vertical distance of the outlier to the core trend has a significant influence on the visual validation of linear trends. Our further analyses indicate that the visual validation of linear trends is highly individual and cognitive once outliers are involved. Further research is required on possible cognitive biases involved in visual validation tasks.

Our work advances our understanding of visual model validation for linear trends in scatterplots containing outliers, and holds practical implications for designers of machine learning applications and visual analytics systems to consider. By systematically evaluating how humans perceive linear relationships in the presence of anomalous data points, our findings inform the design of interfaces and visualization tools that support more accurate human interpretation of model fit, particularly in noisy or imperfect data sets. For practitioners building systems that rely on human-in-the-loop decision-making, such as interactive machine learning platforms or exploratory data analysis tools, accounting for perceptual biases introduced by outliers is critical. Incorporating these insights can improve user trust, prevent misinterpretation of model behavior, and lead to more effective human–AI collaboration.

**Acknowledgments**

The authors would like to thank all study participants. This paper is a result of Dagstuhl Seminar 22331 "Visualization and Decision Making Design Under Uncertainty". This work has been partially supported by the BMBF WarmWorld Project, the SANE Project, and the Risk-Principe Project. This work has also been funded in part by NSF Awards 2007436, 1452977, and 2118201.

# Chapter 4

# Conclusion, Discussion, and Future Directions

This dissertation explored two important yet under-addressed challenges in visual data analysis: the effective communication of large value ranges and the visual validation of regression models. Both topics intersect key phases of the visual analytics pipeline and emphasize the essential role of visualization in supporting human-centered data interpretation.

In addressing the first challenge — visualizing data containing large value ranges — this work introduced novel, scalable visualization techniques that explicitly separate mantissa and exponent components. These techniques, including the *order of magnitude colors* scheme, *order of magnitude line* chart, the *order of magnitude horizon* graph, and the *height-stack line chart*, demonstrated significant improvements in users' ability to interpret complex, time-dependent data. Through extensive user studies, these designs consistently outperformed traditional linear and logarithmic approaches across a range of analytical tasks, such as identification, discrimination, and estimation. The results underline the importance of adapted visual encodings, particularly color, in enhancing perceptual clarity when dealing with multi-order data variability.

The second research focus — visual validation of regression models — expanded the understanding of how users assess model fit through visual means. By empirically studying perceptual biases and user strategies in judging visualized statistical models, this work uncovered critical factors that affect the accuracy and confidence of model validation. A comparative analysis also revealed that visual validation tasks are generally more challenging for users than visual estimation tasks, especially as model complexity increases. These findings inform the development of more interpretable and trustworthy visual analytics systems, particularly in contexts in which explainability and model transparency are paramount.

Together, these contributions advance both the theoretical foundations and practical methodologies of visual data analysis. The proposed visual designs and empirical findings are domain-agnostic, supporting a wide range of applications from meteorology to finance, and from machine learning to public data reporting. By bridging gaps between data representation, user perception, and cognition, this dissertation offers design principles and knowledge about humans' validation abilities that can inform future developments in visualization, explainable AI, and interactive analytics.

The followings sections discuss the implications of this thesis' key findings in detail and outline possible directions for future research.

## 4.1 Visual Communication of Large Value Ranges

Chapter 2 investigated how data with large value ranges can be visualized more effectively, with a focus on time-dependent data. This part of the dissertation was guided by research question **RQ1.1**: *Do visualization designs specially developed for large value ranges in time-dependent data improve the readability and comparability of such data?*

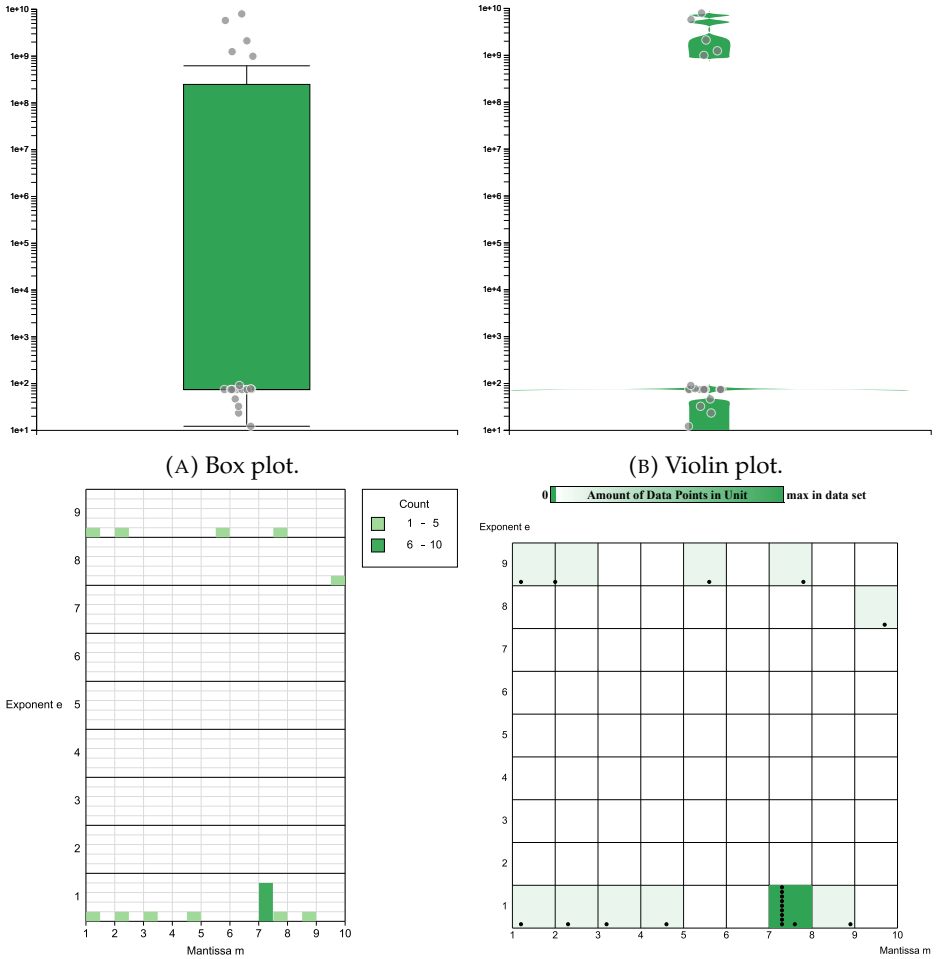
The results of the conducted user studies clearly indicate that visualization designs developed specifically for large value ranges significantly outperform existing state-of-the-art techniques. These improvements are evident across various analytical tasks common in time-series data analysis, including reading, comparison, and estimation tasks. The tailored designs led to measurable improvements in accuracy, response times, and user confidence. These findings provide a clear and affirmative answer to RQ1.1, demonstrating that specialized designs do indeed enhance the readability and comparability of time-dependent data with large value ranges.

Furthermore, a key insight from this work concerns research question **RQ1.2**: *How does the visual mapping to color contribute to the readability of magnitude variations in large value range visualizations?* The results show that color plays a central role as a visual variable in encoding large value differences. Specifically, the use of color supports perceptual differentiation of magnitude variations through "banding effects" — i.e., the utilization of a distinct hue for each exponent creates discernible borders between orders of magnitude. The high performance of the OMC color scale and the new color-based designs confirms that color mapping enhances the effectiveness and intuitiveness of complex visualizations dealing with wide value distributions, thus providing a clear answer to RQ1.2.

Moreover, the benefits of separating mantissa and exponent when representing large values are confirmed. This approach is becoming the common method for encoding magnitudes, but also challenges established graphical perception heuristics. Since a single quantitative value is split across two visual components, users must mentally reconstruct the original value to perform low-level tasks such as reading or comparison. This decomposition increases cognitive load and imposes new design requirements. Potential solutions for these new requirements are provided in this dissertation. While the mantissa-exponent separation has proven effective, future work could explore alternative representations that may offer better perceptual integration without decomposition.

The presented research also uncovered significant patterns in user behavior: as task complexity increases — e.g., when interpreting multiple time-series data sets — users tend to use approximation strategies when facing large value ranges and rely on aspects they already know. This indicates that the more complex the task, the more crucial it becomes for visualizations to be intuitively designed to reduce cognitive overhead. However, not all tasks benefit equally from LVR visualization techniques. Especially for trend detection, LVR-specific designs underperform compared to more traditional representations, most likely due to the non-linear scalings. This highlights the importance of task-specific visualization design.

Looking forward, the work in this dissertation opens several promising **directions for future research on the visual communication of large value ranges**, building upon the existing contributions. Further work remains across all components of the visual analysis process, including data, encoding, interaction and tasks.




(A) Box plot.

(B) Violin plot.

(C) Order of magnitude histogram.

(D) Order of magnitude matrix.

FIGURE 4.1: Example (univariate) data distribution containing large value ranges visualized with the standard designs box plot [184] and violin plot [92] and two prototypes for applying the proven method of separating mantissa and exponent in different ways (histogram versus matrix). The two design proposals were developed as part of a master thesis under this author's supervision [109].

**LVR in other data types** One option for future research is the extension of LVR visualization techniques to additional data types.

The techniques developed and evaluated in this dissertation were constrained to data sets comprising positive values, excluding bipolar data sets that include both positive and negative ranges. The visualization of such data sets presents distinct challenges, particularly in terms of color encoding and scale construction, and will necessitate the development of new visual strategies. Additionally, the data used was centered around base-10 numerical representations. Alternative numerical bases, such as binary (base-2) or the natural base $e$, introduce different exponent distributions, which may require alternative design considerations.

Beyond numerical variations, entirely different data types could be examined. For example, graph data structures pose unique challenges, as both nodes and edges can span large value ranges in their weights. Similarly, dynamic data sets — where data values evolve continuously over time — present perceptual challenges due to the combination of temporal variability and magnitude diversity.

Furthermore, conventional visualization techniques are often inadequate for representing univariate data distributions containing large value ranges (see Figure 4.1). In this context, the prototypes shown in Figure 4.1c and Figure 4.1d illustrate a promising approach: both apply the established principle of separating mantissa and exponent to univariate distributions. The designs display the frequency of mantissa values for each order of magnitude and were developed as part of a master thesis under this author's supervision [109]. However, other techniques — including coloring and binning methods [91] — also need to be evaluated and compared in the context of univariate value distributions.

**Optimization of the visual encoding of LVR** The order of magnitude line chart (OML) design introduced in this work presents a novel scale that combines logarithmic segmentation of magnitudes with linear scaling within each magnitude. While the scaling approach is central to the visual encoding, axis labeling is further important in user perception. Throughout the development process, it was experimented with various labeling strategies — for example, standard logarithmic notation (e.g., "$10^3$"), multiplicative annotations placed within magnitude bands (e.g., "×1000"), or numeric labels positioned at magnitude boundaries (e.g., "1,000"). However, these alternatives were not formally evaluated in a user study. Consequently, a systematic investigation into the perceptual and interpretative effects of different axis scaling and labeling strategies remains important for future research.

Additionally, the optimization of color encoding continues to be a significant area for further development. The OMC scale proposed a "nested" color scheme specifically designed for representing large value ranges. While this dissertation focused on establishing the conceptual foundation of such schemes, the selection and arrangement of specific colors were not thoroughly explored. Future work should aim to refine both the base color palettes and the internal structure of OMC schemes — both globally and within individual magnitudes. Potential optimization goals include improved perceptual smoothness, enhanced uniformity, and increased accessibility, particularly for users with color vision deficiencies.

**Interaction techniques for LVR** This research on the visualization of large value ranges focused on static chart design without considering interaction techniques. Thus, interaction design for data with large value ranges remains underexplored and presents unique challenges. A key difficulty arises from the two-part representation introduced by separating mantissa and exponent. This decomposition complicates traditional interactions such as filtering or brushing, as they can be performed on both a local (i.e., mantissa) or global (i.e., exponent) level. Thus, designing intuitive and efficient interaction techniques that reconcile this separation is a promising area for future work.

Moreover, interactive value range selection becomes particularly challenging in the context of large value distributions. Traditional sliders or range selectors often lack the resolution or granularity to support precise selection across multiple orders of magnitude. New approaches are needed to support both coarse-grained navigation across magnitudes and fine-grained control within them. This functionality could be further enhanced by the implementation of adaptive color coding, which would dynamically adjust based on the selected value range [189].

Another example is the investigation of large value ranges in long time-series, in which local details go unnoticed. Thus, a degree-of-interest (DOI) function for large

value ranges would need to be defined. Moreover, interaction techniques could help users navigate and interpret extended data sets more efficiently [191].

In general, interactivity for LVR data should aim to support a seamless exploration experience, enabling users to flexibly navigate between scales, focus on specific value regions, and compare magnitudes without cognitive overload.

**LVR with high-level tasks** The conducted studies primarily evaluated the visual designs on low-level perceptual tasks, such as value identification, comparison, and difference estimation. While these tasks are foundational for understanding how users perceive and interpret visual encodings for large value ranges, they represent only the lower levels of the analytical task hierarchy [31]. The impact of the proposed visualization techniques on higher-level tasks — such as decision-making or data storytelling — remains an open and important question [64]. These complex tasks involve contextual reasoning, cognitive biases, and narrative framing, which may interact with the visual representation of large value ranges in non-obvious ways. Future research should investigate how LVR visualizations support or hinder such higher-level analytical activities, ideally through applied studies in realistic settings or domain-specific use cases. Understanding the cognitive and communicative implications of these techniques is essential for evaluating their effectiveness beyond isolated perceptual performance.

In conclusion, this thesis demonstrates that effectively communicating large value ranges requires a careful balance between perceptual principles, task demands, and design innovation. The proposed visualization approaches and empirical findings provide a strong foundation for future work aimed at making large value ranges more accessible and interpretable.

## 4.2 Visual Validation of Regression Models

Through a series of human-subject studies, Chapter 3 examined how individuals visually perceive and evaluate regression models. Participants' abilities to visually validate predrawn regression lines and to independently estimate regression trends were compared across varying levels of model complexity, ranging from a baseline model (i.e., the average value) to data exhibiting linear trends and outliers. The findings revealed a consistent — and somewhat counterintuitive — pattern: Participants were generally more accurate and less biased when estimating regression lines themselves, compared to when validating lines presented to them. This suggests that visual estimation may be a more reliable mode of interaction than visual validation in the context of regression analysis.

The findings directly address the research questions posed in this thesis. Regarding **RQ2.1** — *Are individuals able to perform visual validation consistently and without bias for (linear) regression models?* — the results reveal a nuanced picture. While participants demonstrated consistency in their individual validation judgments, their decisions were also systematically biased. This challenges the common assumption that visual validation is a relatively straightforward task due to its binary accept-or-reject nature. In practice, the results suggest the opposite: Visual validation is highly susceptible to perceptual distortions (e.g., the "too steep" bias) and interference from specific data characteristics, such as the presence of outliers. Participants often relied on visual cues — such as the number of points above or below a line or the shape and symmetry of the core distribution — which introduced systematic errors, especially as the visual complexity of the data increased.

Regarding **RQ2.2** — *How does performance in visual model validation relate to the accuracy of visual model estimation (in scatterplots)?* — the findings consistently show that participants were more accurate (i.e., closer to the statistical correct solution) when estimating a trend themselves than when validating one that was shown to them. This difference became more pronounced as the complexity of the model or data increased. In other words, while estimation performance remained relatively stable across varying levels of complexity, validation performance exhibited a decline.

One plausible explanation for these results, based on participants' feedback, is that visual validation may involve an implicit estimation process: Individuals may mentally estimate a trend first, and then judge the validity of the shown trend in relation to their own internal model. If they allow for a margin of error in this process, they may end up accepting lines they would not have drawn themselves. This hypothesis could be tested in future research by asking participants to indicate a range or area in which they believe the true trend lies, rather than estimating an exact trend line.

By systematically increasing the model complexity — starting with the average value and advancing to linear trends that incorporate outliers — the work in this dissertation establishes the foundation for research of visual model validation and creates significant **opportunities for future research**.

**Model type and complexity** The studies in this dissertation explored the visual validation of linear regression models. In the future, more complex or nonlinear model types such as polynomial, exponential, or logistic regression could be investigated. Extending the research to a broader range of model types — especially those used in machine learning applications, such as clustering or classifications [84] — represents an important direction for the further work. These models often exhibit nonlinearity and increased

structural complexity, which pose distinct challenges for intuitive and accurate visual interpretation.

The tasks examined in this work primarily focused on validating individual model parameters, such as slope and intercept. A promising complementary direction involves the visual validation of the model type itself — that is, helping users determine whether the assumed model form appropriately reflects the structure of the data. Supporting such judgments visually could be particularly valuable in exploratory modeling and diagnostic contexts, where misalignment between model and data may not be immediately evident through numerical metrics alone.

Additionally, all experimental studies in this dissertation relied on synthetic data sets. While synthetic data allows for precise control over noise, structure, and confounding variables, it lacks the semantic richness and contextual relevance of real-world data. In applied settings, factors such as domain knowledge, uncertainty, and decision-making may significantly influence how users interpret and validate model outputs. Future research should therefore incorporate real-world data sets to evaluate how these additional cognitive and contextual factors affect visual validation processes.

Furthermore, a deeper investigation into the role of visual complexity could provide valuable insight into users' interpretive performance. A formal definition of visual complexity [157] and its impact on human judgment could improve the understanding of visual validation processes. Such insights may inform the development of more effective visual encoding strategies and support the construction of predictive models of user error, thereby enhancing the robustness of model validation workflows.

**Design guidelines for model visualizations** This thesis evaluated standard regression visualization designs and uncovered several perceptual heuristics and biases that influence how users validate models visually. While formal design guidelines were not formulated, the empirical insights gained provide a strong foundation for developing such principles. Future research should build on these findings to create visualization techniques that intentionally leverage users' perceptual strategies — either by mitigating known biases or by designing in alignment with intuitive judgment patterns. These approaches could be applied both to the visualization of the underlying data, as explored by Liu et al. [116] in the context of visual estimation, and to the visual communication of model outputs.

As an initial step in this direction, a prototype concept for a novel model visualization design was developed. The idea of this approach is to imagine the regression line as a balancing scale, where the goal is to visually assess whether the positive and negative residuals (i.e., the distances of points above and below the line) are in balance (see Figure 4.2). This method aims to make residual symmetry more intuitive and perceptual. However, the concept is still in its early stages and raises several open design questions, such as the parametrization of the encoding.

To further support user judgment, interactive interfaces could incorporate adaptive feedback or confidence cues, enabling users to reflect on their own uncertainty and improve validation accuracy [37, 173]. Such elements could guide attention to regions of interest, flag areas of low confidence, or suggest alternative interpretations based on model uncertainty.

**Trust in visual model validation** Another important area for future research is the role of trust in visual model validation. In the conducted studies in this dissertation, a discrepancy between users' perceived model fit and the actual statistical accuracy of the

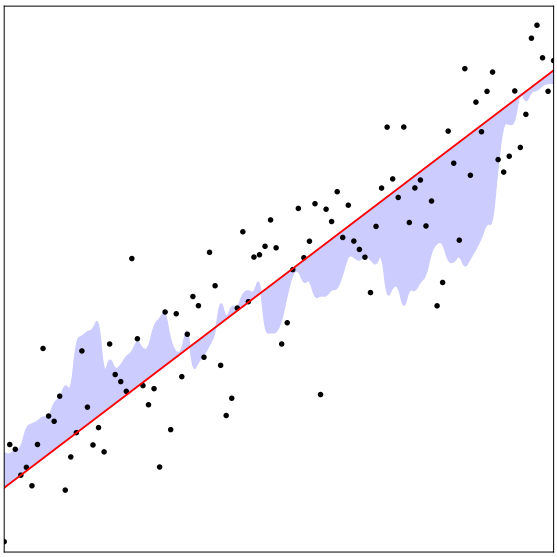


FIGURE 4.2: Illustration of the concept of an "error scales" design for visualizing OLS regression results: The residuals summarized for a rolling window are depicted as areas above and below the line. The total area corresponds to the sum of squared errors. Therefore, a line is the best linear fit to the data, if no adjustment could reduce the area. In this example, a reduction in the slope would decrease the area, suggesting that the current line does not yet minimize the sum of squared errors. Thus, the shown line is too steep.

model was observed. This gap suggests that users' judgments may not only be shaped by perceptual cues, but also by their level of trust in the visualization, the data, or the model itself. Understanding how trust influences trend assessments is particularly critical in high-stakes domains — such as healthcare, finance, or policy — where decisions based on visual interpretations can have significant consequences. Trust may lead users to accept visually plausible but incorrect models, or conversely, to reject valid ones that appear unfamiliar or counterintuitive. Future research should investigate the psychological and contextual factors that shape trust in visualizations, as well as how trust interacts with uncertainty, visual design, and user expertise. Such work could inform the development of trust-aware visualization techniques.

**Visual validation ability of AI** With the growing integration of artificial intelligence (AI) into data analysis workflows, it is becoming increasingly relevant to investigate the visual validation capabilities of AI systems. Specifically, it would be valuable to examine how well AI models can assess regression quality based on visual representations compared to human performance. Such research could shed light on whether AI can replicate or even surpass human abilities in tasks involving perceptual estimation and model validation, as studied in this dissertation. A direct comparison between AI-driven visual validation and human judgment, particularly using controlled image-based inputs similar to those used the experiments of this dissertation, would offer deeper insights into the role AI might play in supporting or automating aspects of visual analysis.

By addressing these areas, future work can contribute to a deeper understanding of human–model interaction and help creating more intuitive, trustworthy, and effective tools for visual data analysis.

# Bibliography


[1] R. Adhikari and R. Agrawal. A combination of artificial neural network and random walk models for financial time series forecasting. *Neural Computing and Applications*, 24:1441–1449, 2014.

[2] M. Adnan, M. Just, and L. Baillie. Investigating time series visualisations to improve the user experience. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 5444–5455. ACM, New York, 2016. doi: 10.1145/2858036.2858300

[3] W. Aigner, C. Kainz, R. Ma, and S. Miksch. Bertin was right: An empirical evaluation of indexing to compare multivariate time-series data using line plots. *Computer Graphics Forum*, 30(1):215–228, 2011. doi: 10.1111/j.1467-8659.2010.01845.x

[4] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *IEEE Transactions on Visualization and Computer Graphics*, 14(1):47–60, 2008. doi: 10.1109/TVCG.2007.70415

[5] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer, London, 2011.

[6] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer, 2023. doi: 10.1007/978-1-4471-7527-8

[7] W. Aigner, A. Rind, and S. Hoffmann. Comparative evaluation of an interactive time-series visualization that combines quantitative data with qualitative abstractions. *Computer Graphics Forum*, 31(3pt2):995–1004, 2012. doi: 10.1111/j.1467-8659.2012.03092.x

[8] D. Albers, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 551—560. ACM, 2014. doi: 10.1145/2556288.2557200

[9] D. Altman, D. Machin, T. Bryant, and M. Gardner. *Statistics with confidence: confidence intervals and statistical guidelines*. BMJ Books, London, 2000.

[10] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*, pp. 111–117, 2005. doi: 10.1109/INFVIS.2005.1532136

[11] N. Andrienko, G. Andrienko, L. Adilova, and S. Wrobel. Visual analytics for human-centered machine learning. *IEEE Computer Graphics and Applications*, 42(1):123–133, 2022. doi: 10.1109/MCG.2021.3130314

[12] N. Andrienko, T. Lammarsch, G. Andrienko, G. Fuchs, D. Keim, S. Miksch, and A. Rind. Viewing visual analytics as model building. *Computer Graphics Forum*, 37(6):275–299, 2018. doi: 10.1111/cgf.13324

[13] S. Ansari, A. B. Nassif, S. Mahmoud, S. Majzoub, E. Almajali, A. Jarndal, T. Bonny, K. A. Alnajjar, and A. Hussain. Impact of outliers on regression and classification models: An empirical analysis. In *2024 17th International Conference on Development in eSystem Engineering (DeSE)*, pp. 211–218, 2024. doi: 10.1109/DeSE63988.2024.10912020

[14] D. R. Baldo, M. S. Regio, and I. H. Manssour. Visual analytics for monitoring credit scoring models. *Information Visualization*, 22(4):340–357, 2023. doi: 10.1177/14738716231180803

[15] L. Bartram and M. C. Stone. Whisper, don't scream: Grids and transparency. *IEEE Transactions on Visualization and Computer Graphics*, 17(10):1444–1458, 2011. doi: 10.1109/TVCG.2010.237

[16] S. Bateman, R. L. Mandryk, C. Gutwin, A. Genest, D. McDine, and C. Brooks. Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2573–2582. ACM, New York, 2010. doi: 10.1145/1753326.1753716

[17] K. Batziakoudi, F. Cabric, S. Rey, and J.-D. Fekete. A Design Space for Static Visualizations with Several Orders of Magnitude. In *EuroVis 2024 - Posters*. Eurographics, 2024. doi: 10.2312/evp.20241076

[18] K. Batziakoudi, F. Cabric, S. Rey, and J.-D. Fekete. Lost in magnitudes: Exploring visualization designs for large value ranges. In *CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 2025. doi: 10.1145/3706598.3713487

[19] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, Madison, 1983.

[20] V. Bewick, L. Cheek, and J. Ball. Statistics review 7: Correlation and regression. *Critical care*, 7(6):451–459, 2003. doi: 10.1186/cc2401

[21] C. R. Blyth. On simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67(338):364–366, 1972. doi: 10.1080/01621459.1972.10482387

[22] M. Bögl, W. Aigner, P. Filzmoser, T. Lammarsch, S. Miksch, and A. Rind. Visual analytics for model selection in time series analysis. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2237–2246, 2013. doi: 10.1109/TVCG.2013.222

[23] R. Borgo, J. Dearden, and M. W. Jones. Order of magnitude markers: An empirical study on large magnitude number detection. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2261–2270, 2014. doi: 10.1109/TVCG.2014.2346428

[24] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister. What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2306–2315, 2013. doi: 10.1109/TVCG.2013.234

[25] D. Braun, R. Borgo, M. Sondag, and T. von Landesberger. Reclaiming the horizon: Novel visualization designs for time-series data with large value ranges. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1161–1171, 2024. doi: 10.1109/TVCG.2023.3326576

[26] D. Braun, R. Borgo, M. Sondag, and T. von Landesberger. Design and evaluation of visualizations for large value ranges in multiple time-series. *Information Visualization*, 2025. doi: 10.1177/14738716251349501

[27] D. Braun, R. Chang, M. Gleicher, and T. von Landesberger. Beware of validation by eye: Visual validation of linear trends in scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):787–797, 2025. doi: 10.1109/TVCG.2024.3456305

[28] D. Braun, R. Chang, M. Gleicher, and T. von Landesberger. Visual validation of linear trends in scatterplots amid outliers. *In preparation for submission to CHI conference*, 2026.

[29] D. Braun, K. Ebell, V. Schemann, L. Pelchmann, S. Crewell, R. Borgo, and T. von Landesberger. Color coding of large value ranges applied to meteorological data. In *2022 IEEE Visualization and Visual Analytics (VIS)*, pp. 125–129, 2022. doi: 10.1109/VIS54862.2022.00034

[30] D. Braun, A. Suh, R. Chang, M. Gleicher, and T. von Landesberger. Visual validation versus visual estimation: A study on the average value in scatterplots. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 181–185, 2023. doi: 10.1109/VIS54172.2023.00045

[31] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2376–2385, 2013. doi: 10.1109/TVCG.2013.124

[32] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 83–92, 2012. doi: 10.1109/VAST.2012.6400486

[33] 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 Transactions on Visualization and Computer Graphics*, 27(2):1634–1643, 2021. doi: 10.1109/TVCG.2020.3030404

[34] J. Buchmüller, D. Jäckle, E. Cakmak, U. Brandes, and D. A. Keim. Motionrugs: Visualizing collective trends in space and time. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):76–86, 2019. doi: 10.1109/TVCG.2018.2865049

[35] S. Card and J. Mackinlay. The structure of the information visualization design space. In *Proceedings of VIZ '97: Visualization Conference, Information Visualization Symposium and Parallel Rendering Symposium*, pp. 92–99, 1997. doi: 10.1109/INFVIS.1997.636792

[36] D. Cashman, S. R. Humayoun, F. Heimerl, K. Park, S. Das, J. Thompson, B. Saket, A. Mosca, J. Stasko, A. Endert, et al. A user-based visual analytics workflow for exploratory model analysis. *Computer Graphics Forum*, 38(3):185–199, 2019. doi: 10.1111/cgf.13681

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

[38] G. Charness, U. Gneezy, and M. A. Kuhn. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1):1–8, 2012. doi: 10.1016/j.jebo.2011.08.009

[39] A. Chatzimparmpas, K. Kucher, and A. Kerren. Visualization for trust in machine learning revisited: The state of the field in 2023. *IEEE Computer Graphics and Applications*, 44(3):99–113, 2024. doi: 10.1109/MCG.2024.3360881

[40] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233, 2020. doi: 10.1177/1473871620904671

[41] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. *Computer Graphics Forum*, 39(3):713–756, 2020. doi: 10.1111/cgf.14034

[42] M. Chegini, L. Shao, R. Gregor, D. J. Lehmann, K. Andrews, and T. Schreck. Interactive visual exploration of local patterns in large scatterplot spaces. *Computer Graphics Forum*, 37(3):99–109, 2018. doi: 10.1111/cgf.13404

[43] I. K. Choi, N. K. Raveendranath, J. Westerfield, and K. Reda. Visual (dis)confirmation: Validating models and hypotheses with visualizations. In *2019 23rd International Conference in Information Visualization – Part II*, pp. 116–121, 2019. doi: 10.1109/IV-2.2019.00032

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

[45] L. Ciccione, G. Dehaene, and S. Dehaene. Outlier detection and rejection in scatterplots: Do outliers influence intuitive statistical judgments? *Journal of Experimental Psychology: Human Perception and Performance*, 49:129–144, 2023. doi: 10.1037/xhp0001065

[46] L. Ciccione and S. Dehaene. Can humans perform mental regression on a graph? accuracy and bias in the perception of scatterplots. *Cognitive Psychology*, 128, 2021. doi: 10.1016/j.cogpsych.2021.101406

[47] L. Ciccione, M. Sablé-Meyer, and S. Dehaene. Analyzing the misperception of exponential growth in graphs. *Cognition*, 225:105112, 2022. doi: 10.1016/j.cognition.2022.105112

[48] J. H. Clark. The ishihara test for color blindness. *American Journal of Physiological Optics*, 5:269–276, 1924.

[49] K. A. Cook and J. J. Thomas. *Illuminating the path: The research and development agenda for visual analytics*. IEEE, 2005.

[50] M. Correll. Ethical dimensions of visualization research. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–13, 2019.

[51] M. Correll, D. Albers, S. Franconeri, and M. Gleicher. Comparing averages in time series data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1095—1104. ACM, 2012. doi: 10.1145/2207676.2208556

[52] M. Correll and J. Heer. Regression by eye: Estimating trends in bivariate visualizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1387—-1396. ACM, New York, 2017. doi: 10.1145/3025453.3025922

[53] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger. Looks good to me: Visualizations as sanity checks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):830–839, 2019. doi: 10.1109/TVCG.2018.2864907

[54] J. Crouch. Mapping u.s. climate trends. `https://www.climate.gov/news-features/blogs/beyond-data/mapping-us-climate-trends`, 2017.

[55] D. Dansana, R. Kumar, J. Das Adhikari, M. Mohapatra, R. Sharma, I. Priyadarshini, and D.-N. Le. Global forecasting confirmed and fatal cases of covid-19 outbreak using autoregressive integrated moving average model. *Frontiers in Public Health*, 8:580327:1–580327:11, 2020. doi: 10.3389/fpubh.2020.580327

[56] A. Dasgupta, J.-Y. Lee, R. Wilson, R. A. Lafrance, N. Cramer, K. Cook, and S. Payne. Familiarity vs trust: A comparative study of domain scientists' trust in visual analytics and conventional analysis methods. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):271–280, 2017. doi: 10.1109/TVCG.2016.2598544

[57] R. M. Dawes. The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7):571–582, 1979. doi: 10.1037/0003-066X.34.7.571

[58] R. M. Dawes and B. Corrigan. Linear models in decision making. *Psychological Bulletin*, 81(2):95–106, 1974. doi: 10.1037/h0037613

[59] I. Demir, M. Jarema, and R. Westermann. Visualizing the central tendency of ensembles of shapes. In *SIGGRAPH ASIA 2016 Symposium on Visualization*, pp. 3:1–3:8. ACM, New York, 2016. doi: 10.1145/3002151.3002165

[60] B. der Finanzen. Bundeshaushalt digital. `https://www.bundeshaushalt.de/DE/Bundeshaushalt-digital/bundeshaushalt-digital.html`. Location: Berlin, DE. Accessed: 2025-05-09.

[61] A. Diehl, A. Abdul-Rahman, M. El-Assady, B. Bach, D. Keim, and M. Chen. Visguides: A forum for discussing visualization guidelines. In *EuroVis 2018 - Short Papers*, pp. 61–65. The Eurographics Association, Eindhoven, 2018. doi: 10.2312/eurovisshort.20181079

[62] S. A. A. Dilawer and S. R. Humayoun. A visual analytics tool to explore multi-classification model with high number of classes. In *Companion Proceedings of the 16th ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, EICS '24 Companion, 3 pages, p. 84–86. ACM, 2024. doi: 10.1145/3660515.3662833

[63] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic. A task-based taxonomy of cognitive biases for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 26(2):1413–1432, 2020. doi: 10.1109/TVCG.2018.2872577

[64] E. Dimara and J. Stasko. A critical reflection on visualization research: Where do decision making tasks hide? *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1128–1138, 2022. doi: 10.1109/TVCG.2021.3114813

[65] C. Dubé. Central tendency representation and exemplar matching in visual short-term memory. *Memory & Cognition*, 47:589–602, 2019. doi: 10.3758/s13421-019-00900-0

[66] H. Elhamdadi, A. Gaba, Y.-S. Kim, and C. Xiong. How do we measure trust in visual data communication? In *2022 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pp. 85–92. IEEE, New York, 2022. doi: 10.1109/BELIV57783.2022.00014

[67] A. Endert, P. Fiaux, and C. North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 473–482. ACM, 2012.

[68] Y. Fang, H. Xu, and J. Jiang. A survey of time series data visualization research. *IOP Conference Series: Materials Science and Engineering*, 782(2):022013:1–022013:10, 2020. doi: 10.1088/1757-899X/782/2/022013

[69] P. Federico, S. Hoffmann, A. Rind, W. Aigner, and S. Miksch. Qualizon graphs: Space-efficient time-series visualization with qualitative abstractions. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*, pp. 273—280. ACM, New York, 2014. doi: 10.1145/2598153.2598172

[70] S. Few. Grid lines in graphs are rarely useful. *DM Review*, 15(2):46:1–46:4, 2005.

[71] S. Few. Time on the horizon. *Visual Business Intelligence Newsletter*, 2008.

[72] S. Few. The chartjunk debate. *Visual Business Intelligence Newsletter*, 2011.

[73] A. Filipowicz, S. Carter, N. Bravo, R. Iliev, S. Hakimi, D. A. Shamma, K. Lyons, C. Hogan, and C. Wu. Visual elements and cognitive biases influence interpretations of trends in scatter plots, 2023.

[74] Finnish Meteorological Institute. Cloudnet data portal. `https://cloudnet.fmi.fi/search/visualizations`.

[75] M. Franke., M. Knabben., J. Lang., S. Koch., and T. Blascheck. A comparative study of visualizations for multiple time series. In *Proceedings of VISIGRAPP*, pp. 103–112. SciTePress, 2022. doi: 10.5220/0010761700003124

[76] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, p. 3237–3246. ACM, 2013. doi: 10.1145/2470654.2466443

[77] J. Fuchs, P. Isenberg, A. Bezerianos, M. Miller, and D. Keim. EduClust - A Visualization Application for Teaching Clustering Algorithms. In *Eurographics 2019 - Education Papers*. Eurographics, 2019. doi: 10.2312/eged.20191023

[78] M. Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, 2018. doi: 10.1109/TVCG.2017.2744199

[79] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011. doi: 10.1177/1473871611416549

[80] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2316–2325, 2013. doi: 10.1109/TVCG.2013.183

[81] A. Gogolou, T. Tsandilas, T. Palpanas, and A. Bezerianos. Comparing similarity perception in time series visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):523–533, 2019. doi: 10.1109/TVCG.2018.2865077

[82] M. Gonzalez, C. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008. doi: 10.1038/nature06958

[83] I. M. Gołbiowska and A. Çöltekin. Rainbow dash: Intuitiveness, interpretability and memorability of the rainbow color scheme in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(7):2722–2733, 2022. doi: 10.1109/TVCG.2020.3035823

[84] N. Grossmann, J. Bernard, M. Sedlmair, and M. Waldner. Does the layout really matter? a study on visual model accuracy estimation. In *2021 IEEE Visualization Conference (VIS)*, pp. 61–65, 2021. doi: 10.1109/VIS49827.2021.9623326

[85] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014. doi: 10.1109/TVCG.2014.2346979

[86] M. Harrower and C. A. Brewer. Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. doi: 10.1179/000870403235002042

[87] S. Havre, B. Hetzler, and L. Nowell. Themeriver: visualizing theme changes over time. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000.*, pp. 115–123. IEEE, New York, 2000. doi: 10.1109/INFVIS.2000.885098

[88] E. Hawkins. Show your stripes. `https://showyourstripes.info`, 2018. Accessed: 01.02.2023.

[89] Y. He and H. Li. Optimal layout of stacked graph for visualizing multidimensional financial time series data. *Information Visualization*, 21(1):63–73, 2022. doi: 10.1177/14738716211045005

[90] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1303—-1312. ACM, 2009. doi: 10.1145/1518701.1518897

[91] A. Heim, A. Gall, M. Waldner, E. Gröller, and C. Heinzl. Accustripes: Visual exploration and comparison of univariate data distributions using color and binning. *Computers & Graphics*, 119:103906, 2024. doi: 10.1016/j.cag.2024.103906

[92] J. L. Hintze and R. D. Nelson. Violin plots: a box plot-density trace synergism. *The American Statistician*, 52(2):181–184, 1998. doi: 10.1080/00031305.1998.10480559

[93] M. Hlawatsch, F. Sadlo, M. Burch, and D. Weiskopf. Scale-Stack Bar Charts. *Computer Graphics Forum*, 32(3pt2):181–190, 2013. doi: 10.1111/cgf.12105

[94] M. Höhn, M. Wunderlich, K. Ballweg, and T. von Landesberger. Width-scale bar charts for data with large value range. *EuroVis 2020 - Short Papers*, 2020. doi: 10.2312/evs.20201056

[95] M.-H. Hong, J. K. Witt, and D. A. Szafir. The weighted average illusion: Biases in perceived mean position in scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):987–997, 2022. doi: 10.1109/TVCG.2021.3114783

[96] J. Hullman and A. Gelman. Designing for interactive exploratory data analysis requires theories of graphical inference. *Harvard Data Science Review*, 3(3), 2021. doi: 10.1162/99608f92.3ab8a587

[97] Institute for Geophysics and Meteorology, University of Cologne. Measurments and visualizations of meteorological cloud data. `https://atmos.meteo.uni-koeln.de/~hatpro/dataBrowser/`.

[98] A. Jabbari, R. Blanch, and S. Dupuy-Chessa. Beyond horizon graphs: Space efficient time series visualization with composite visual mapping. In *Proceedings of the 30th Conference on l'Interaction Homme-Machine*, pp. 73—–82. ACM, New York, 2018. doi: 10.1145/3286689.3286694

[99] A. Jabbari, R. Blanch, and S. Dupuy-Chessa. Composite visual mapping for time series visualization. In *2018 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 116–124. IEEE, New York, 2018. doi: 10.1109/PacificVis.2018.00023

[100] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934, 2010. doi: 10.1109/TVCG.2010.162

[101] D. H. Jeong, C. Ziemkiewicz, B. Fisher, W. Ribarsky, and R. Chang. ipca: An interactive system for pca-based visual analytics. *Computer Graphics Forum*, 28(3):767–774, 2009. doi: 10.1111/j.1467-8659.2009.01475.x

[102] A. B. M. Jeuken, P. C. Siegmund, L. C. Heijboer, J. Feichter, and L. Bengtsson. On the potential of assimilating meteorological analyses in a global climate model for the purpose of model validation. *Journal of Geophysical Research: Atmospheres*, 101(D12):16939–16950, 1996. doi: 10.1029/96JD01218

[103] L. Jian, C. Russell, J. Luhmann, and R. M. Skoug. Properties of interplanetary coronal mass ejections at one au during 1995–2004. *Solar Physics*, 239:393–436, 2006. doi: 10.1007/s11207-006-0133-2

[104] H. Kanamori. The energy release in great earthquakes. *Journal of Geophysical Research*, 82(20):2981–2987, 1977. doi: 10.1029/JB082i020p02981

[105] R. M. Karim, O.-H. Kwon, C. Park, and K. Lee. A study of colormaps in network visualization. *Applied Sciences*, 9(20), article no. 4228, 2019. doi: 10.3390/app9204228

[106] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):469–478, 2016. doi: 10.1109/TVCG.2015.2467671

[107] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. *Visual Analytics: Scope and Challenges*, pp. 76–90. Springer, 2008. doi: 10.1007/978-3-540-71080-6_6

[108] N. Kerracher, J. Kennedy, and K. Chalmers. A task taxonomy for temporal graph visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 21(10):1160–1172, 2015. doi: 10.1109/TVCG.2015.2424889

[109] M. Klein. *Visualization of univariate data distributions with large value ranges*. Master Thesis at the University of Cologne, 2024.

[110] M. Krstajic, E. Bertini, and D. Keim. Cloudlines: Compact display of event episodes in multiple time-series. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2432–2439, 2011. doi: 10.1109/TVCG.2011.179

[111] R. M. Kubina, S. A. King, M. Halkowski, S. Quigley, and T. Kettering. Slope identification and decision making: A comparison of linear and ratio graphs. *Behavior Modification*, 47(3):615–643, 2023. doi: 10.1177/01454455221130002

[112] T. H. Le, T. K. Dang, T. T. Dang, and T. N. Luong. Visualizing access logs of a scientific digital library effectively as multiple time series using modified horizon graphs. *2019 International Conference on Advanced Computing and Applications (ACOMP)*, pp. 85–91, 2019. doi: 10.1109/ACOMP.2019.00020

[113] H. Lee, J. Kihm, J. Choo, J. Stasko, and H. Park. ivisclustering: An interactive visual document clustering via topic modeling. *Computer Graphics Forum*, 31(3pt3):1155–1164, 2012. doi: 10.1111/j.1467-8659.2012.03108.x

[114] Y. Li, T. Fujiwara, Y. K. Choi, K. K. Kim, and K.-L. Ma. A visual analytics system for multi-model comparison on clinical data predictions. *Visual Informatics*, 4(2):122–131, 2020. PacificVis 2020 Workshop on Visualization Meets AI. doi: 10.1016/j.visinf.2020.04.005

[115] LimeSurvey Project Team / Carsten Schmitz. Limesurvey: An open source survey tool. `http://www.limesurvey.org`, 2024.

[116] T. Liu, X. Li, C. Bao, M. Correll, C. Tu, O. Deussen, and Y. Wang. Data-driven mark orientation for trend estimation in scatterplots. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, article no. 473, 16 pages. ACM, 2021. doi: 10.1145/3411764.3445751

[117] Y. Lu, R. Garcia, B. Hansen, M. Gleicher, and R. Maciejewski. The state-of-the-art in predictive visual analytics. *Computer Graphics Forum*, 36(3):539–562, 2017. doi: 10.1111/cgf.13210

[118] S. LYi, J. Jo, and J. Seo. Comparative layouts revisited: Design space, guidelines, and future directions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1525–1535, 2021. doi: 10.1109/TVCG.2020.3030419

[119] H. H. Mahbubul Majumder and D. Cook. Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, 108(503):942–956, 2013. doi: 10.1080/01621459.2013.808157

[120] M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri. Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *Journal of medical systems*, 42:1–17, 2018. doi: 10.1007/s10916-018-0940-7

[121] H. Márquez-González, M. G. Miranda-Novales, F. Solórzano-Santos, M. Klunder-Klunder, J. Garduño-Espinoza, and J. F. Méndez-Galván. Covid-19 pandemic: challenges ahead. *Boletín médico del Hospital Infantil de México*, 77(5):242–251, 2020.

[122] J. Matejka and G. Fitzmaurice. Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1290—1294. ACM, New York, 2017. doi: 10.1145/3025453.3025912

[123] E. Mayr, N. Hynek, S. Salisu, and F. Windhager. Trust in information visualization. In *EuroVis Workshop on Trustworthy Visualization (TrustVis)*, pp. 25–29. Eurographics Association, Goslar, 2019. doi: 10.2312/trvis.20191187

[124] C. M. McColeman, L. Harrison, M. Feng, and S. Franconeri. No mark is an island: Precision and category repulsion biases in data reproductions. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1063–1072, 2021. doi: 10.1109/TVCG.2020.3030345

[125] D. N. Menge, A. C. MacPherson, T. A. Bytnerowicz, A. W. Quebbeman, N. B. Schwartz, B. N. Taylor, and A. A. Wolf. Logarithmic scales in ecological data presentation may cause misinterpretation. *Nature Ecology & Evolution*, 2(9):1393–1402, 2018. doi: 10.1038/s41559-018-0610-7

[126] D. Moritz, L. M. Padilla, F. Nguyen, and S. L. Franconeri. Average estimates in line graphs are biased toward areas of higher variability. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):306–315, 2024. doi: 10.1109/TVCG.2023.3326589

[127] V. Morton and D. J. Torgerson. Effect of regression to the mean on decision making in health care. *BMJ (Clinical research ed.)*, 326(7398):1083–1084, 2003. doi: 10.1136/bmj.326.7398.1083

[128] A. Motrenko, V. Strijov, and G.-W. Weber. Sample size determination for logistic regression. *Journal of Computational and Applied Mathematics*, 255:743–752, 2014. doi: 10.1016/j.cam.2013.06.031

[129] T. Munzner. *Visualization Analysis and Design*. A K Peters/CRC Press, Florida, 2014.

[130] T. Mühlbacher and H. Piringer. A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1962–1971, 2013. doi: 10.1109/TVCG.2013.125

[131] P. Nardini, M. Chen, M. Böttinger, G. Scheuermann, and R. Bujack. Automatic improvement of continuous colormaps in euclidean colorspaces. *Computer Graphics Forum*, 40(3):361–373, 2021. doi: 10.1111/cgf.14313

[132] E. Newburger, M. Correll, and N. Elmqvist. Fitting bell curves to data distributions using visualization. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–12, 2022. doi: 10.1109/TVCG.2022.3210763

[133] B. Nibley. Bitcoin price history: 2009 - 2023. https://www.sofi.com/learn/content/bitcoin-price-history/, 2022. Accessed: 27.01.2023.

[134] J. R. Nuñez, C. R. Anderton, and R. S. Renslow. Optimizing colormaps with consideration for color vision deficiency to enable accurate interpretation of scientific data. *PLOS ONE*, 13(7):1–14, 2018. doi: 10.1371/journal.pone.0199239

[135] B. Ondov, N. Jardine, N. Elmqvist, and S. Franconeri. Face to face: Evaluating visual comparison. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):861–871, 2019. doi: 10.1109/TVCG.2018.2864884

[136] B. D. Ondov, F. Yang, M. Kay, N. Elmqvist, and S. Franconeri. Revealing perceptual proxies with adversarial examples. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1073–1083, 2021. doi: 10.1109/TVCG.2020.3030429

[137] B. Oral and A. Boduroglu. Effects of outlier and familiar context in trend-line estimates in scatterplots. *Memory & Cognition*, pp. 1–19, 2024. doi: 10.3758/s13421-024-01646-0

[138] L. Padilla, S. C. Castro, and H. Hosseinpour. Chapter seven - a review of uncertainty visualization errors: Working memory as an explanatory theory. In K. D. Federmeier, ed., *The Psychology of Learning and Motivation*, vol. 74 of *Psychology of Learning and Motivation*, pp. 275–315. Academic Press, 2021. doi: 10.1016/bs.plm.2021.03.001

[139] L. Padilla, R. Fygenson, S. C. Castro, and E. Bertini. Multiple forecast visualizations (mfvs): Trade-offs in trust and performance in multiple covid-19 forecast visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):12–22, 2023. doi: 10.1109/TVCG.2022.3209457

[140] L. Padilla, M. Kay, and J. Hullmann. Uncertainty visualization. In W. Piegorsch, R. Levine, H. Zhang, and T. Lee, eds., *Computational Statistics in Data Science*, pp. 405–421. Wiley, 2022.

[141] C. Perin, F. Vernier, and J.-D. Fekete. Interactive horizon graphs: Improving the compact visualization of multiple time series. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 3217—3226. ACM, 2013. doi: 10.1145/2470654.2466441

[142] D. Peskov, B. Cheng, A. Elgohary, J. Barrow, C. Danescu-Niculescu-Mizil, and J. Boyd-Graber. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3811–3854. ACL, Stroudsburg, 2020. doi: 10.18653/v1/2020.acl-main.353

[143] T. Piketty and E. Saez. Income inequality in the united states, 1913–1998. *The Quarterly Journal of Economics*, 118(1):1–41, 2003. doi: 10.1162/00335530360535135

[144] D. R. Proffitt, M. Bhalla, R. Gossweiler, and J. Midgett. Perceiving geographical slant. *Psychonomic Bulletin & Review*, 2:409–428, 1995. doi: 10.3758/BF03210980

[145] Prolific. Prolific · quickly find research participants you can trust. https://www.prolific.co. Location: London, UK. Accessed: 2024-03-12.

[146] G. J. Quadri and P. Rosen. A survey of perception-based visualization studies by task. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):5026–5048, 2022. doi: 10.1109/TVCG.2021.3098240

[147] R Core Team. R: A language and environment for statistical computing. `https://www.R-project.org/`, 2025.

[148] K. Reda, A. A. Salvi, J. Gray, and M. E. Papka. Color nameability predicts inference accuracy in spatial visualizations. *Computer Graphics Forum*, 40(3):49–60, 2021. doi: 10.1111/cgf.14288

[149] K. Reda and D. A. Szafir. Rainbows revisited: Modeling effective colormap design for graphical inference. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1032–1042, 2021. doi: 10.1109/TVCG.2020.3030439

[150] H. Reijner. The development of the horizon graph. In *Electronic Proceedings of the Vis08 Workshop From Theory to Practice: Design, Vision and Visualization*. IEEE, New York, 2008.

[151] D. Reimann, C. Blech, N. Ram, and R. Gaschler. Visual model fit estimation in scatterplots: Influence of amount and decentering of noise. *IEEE Transactions on Visualization and Computer Graphics*, 27(9):3834–3838, 2021. doi: 10.1109/TVCG.2021.3051853

[152] R. A. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum*, 29(3):1203–1210, 2010. doi: 10.1111/j.1467-8659.2009.01694.x

[153] B. E. Rogowitz and L. A. Treinish. Data visualization: the end of the rainbow. *IEEE Spectrum*, 35(12):52–59, 1998. doi: 10.1109/6.736450

[154] B. E. Rogowitz, L. A. Treinish, and S. Bryson. How not to lie with visualization. *Computers in Physics and IEEE Computational Science & Engineering*, 10(3):268–273, 1996. doi: 10.1063/1.4822401

[155] A. Romano, C. Sotis, G. Dominioni, and S. Guidi. The scale of covid-19 graphs affects understanding, attitudes, and policy preferences. *Health Economics*, 29(11):1482–1494, 2020. doi: 10.1002/hec.4143

[156] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2):154–172, 2016. doi: 10.1080/13875868.2015.1137577

[157] G. Ryan, A. Mosca, R. Chang, and E. Wu. At a glance: Pixel approximate entropy as a measure of line chart complexity. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):872–881, 2019. doi: 10.1109/TVCG.2018.2865264

[158] D. Sacha, M. Sedlmair, L. Zhang, J. Lee, D. Weiskopf, S. North, and D. Keim. Human-centered machine learning through interactive visualization. In *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pp. 641–646, 2016.

[159] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):240–249, 2016. doi: 10.1109/TVCG.2015.2467591

[160] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda. Two-tone pseudo coloring: compact visualization for one-dimensional data. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005*, pp. 173–180. IEEE, New York, 2005. doi: 10.1109/INFVIS.2005.1532144

[161] A. Sarikaya and M. Gleicher. Scatterplots: Tasks, data, and designs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):402–412, 2018. doi: 10.1109/TVCG.2017.2744184

[162] V. Schemann and K. Ebell. Simulation of mixed-phase clouds with the icon large-eddy model in the complex arctic environment around ny-Ålesund. *Atmospheric Chemistry and Physics*, 20(1):475–485, 2020. doi: 10.5194/acp-20-475-2020

[163] P. Schulze-Wollgast, C. Tominski, and H. Schumann. Enhancing visual exploration by appropriate color coding. *Proceedings of International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, pp. 203–210, 2005.

[164] S. Sevi, M. M. Aviña, G. Péloquin-Skulski, E. Heisbourg, P. Vegas, M. Coulombe, V. Arel-Bundock, P. J. Loewen, and A. Blais. Logarithmic versus linear visualizations of covid-19 cases do not affect citizens' support for confinement. *Canadian Journal of Political Science*, 53(2):385–390, 2020. doi: 10.1017/S000842392000030X

[165] L. Shao, A. Mahajan, T. Schreck, and D. J. Lehmann. Interactive regression lens for exploring scatter plots. *Computer Graphics Forum*, 36(3):157–166, 2017. doi: 10.1111/cgf.13176

[166] G. Sharma, W. Wu, and E. N. Dalal. The ciede2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research and Application*, 30(1):21–30, 2005.

[167] J. Sheidin, J. Lanir, and T. Kuflik. A comparative evaluation of techniques for time series visualizations of emotions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, article no. 21, 9 pages. ACM, 2019. doi: 10.1145/3351995.3352054

[168] C. Shi, W. Cui, S. Liu, P. Xu, W. Chen, and H. Qu. Rankexplorer: Visualization of ranking changes in large time series data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2669–2678, 2012. doi: 10.1109/TVCG.2012.253

[169] E. H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951. doi: 10.1111/j.2517-6161.1951.tb00088.x

[170] G. Skok. Spiral strip. *Applied Sciences*, 12(13), article no. 6609, 2022. doi: 10.3390/app12136609

[171] S. Smart, K. Wu, and D. A. Szafir. Color crafting: Automating the construction of designer quality color ramps. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1215–1225, 2020. doi: 10.1109/TVCG.2019.2934284

[172] M. Soprano, K. Roitero, D. La Barbera, D. Ceolin, D. Spina, G. Demartini, and S. Mizzaro. Cognitive biases in fact-checking and their countermeasures: A review. *Information Processing & Management*, 61(3):103672, 2024. doi: 10.1016/j.ipm.2024.103672

[173] F. Sperrle, A. Jeitler, J. Bernard, D. Keim, and M. El-Assady. Co-adaptive visual data analysis and guidance processes. *Computers & Graphics*, 100:93–105, 2021. doi: 10.1016/j.cag.2021.06.016

[174] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1064–1074, 2020. doi: 10.1109/TVCG.2019.2934629

[175] M. Stone and L. Bartram. Alpha, contrast and the perception of visual metadata. In *Color and Imaging Conference*, vol. 16, pp. 355–359. Society of Imaging Science and Technology, Springfield, 2008.

[176] G. Strain, A. J. Stewart, P. Warren, and C. Jay. Adjusting point size to facilitate more accurate correlation perception in scatterplots. In *2023 IEEE Vis X Vision*, pp. 1–5, 2023. doi: 10.1109/VisXVision60716.2023.00006

[177] G. Strain, A. J. Stewart, P. Warren, and C. Jay. The effects of contrast on correlation perception in scatterplots. *International Journal of Human-Computer Studies*, 176, 2023. doi: 10.1016/j.ijhcs.2023.103040

[178] A. Sultan, W. Sałabun, S. Faizi, and M. Ismail. Hesitant fuzzy linear regression model for decision making. *Symmetry*, 13(10), 2021. doi: 10.3390/sym13101846

[179] SurveyCircle Project Team. Research website surveycircle. `https://www.surveycircle.com`, 2025.

[180] D. A. Szafir. Modeling color difference for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):392–401, 2018. doi: 10.1109/TVCG.2017.2744359

[181] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of Vision*, 16(5):11:1–11:19, 2016. doi: 10.1167/16.5.11

[182] Tableau. Add trend lines to a visualization. `https://help.tableau.com/current/pro/desktop/en-us/trendlines_add.htm`. Location: Seattle, US. Accessed: 2024-03-04.

[183] C. Tominski, G. Fuchs, and H. Schumann. Task-driven color coding. *12th International Conference Information Visualisation (IV)*, pp. 373–380, 2008. doi: 10.1109/IV.2008.24

[184] J. W. Tukey. *Exploratory data analysis*. Addison-Wesley, 1977.

[185] E. W. Tyree and J. Long. Forecasting currency exchange rates: Neural networks and the random walk model. In *City University Working Paper, Proc. of the Third International Conference on Artificial Intelligence Applications*, 1995.

[186] W. M. Vagias. Likert-type scale response anchors. *Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson University*, 2006.

[187] E. R. A. Valiati, M. S. Pimenta, and C. M. D. S. Freitas. A taxonomy of tasks for guiding the evaluation of multidimensional visualizations. *BELIV '06: Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pp. 1–6, 2006. doi: 10.1145/1168149.1168169

[188] S. van den Elzen and J. J. van Wijk. Baobabview: Interactive construction and analysis of decision trees. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 151–160, 2011. doi: 10.1109/VAST.2011.6102453

[189] N. Waldin, M. Waldner, M. Le Muzic, E. Gröller, D. S. Goodsell, L. Autin, A. J. Olson, and I. Viola. Cuttlefish: Color mapping for dynamic multi-scale visualizations. *Computer Graphics Forum*, 38(6):150–164, 2019. doi: 10.1111/cgf.13611

[190] M. Waldner, A. Diehl, D. Gračanin, R. Splechtna, C. Delrieux, and K. Matković. A comparison of radial and linear charts for visualizing daily patterns. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1033–1042, 2020. doi: 10.1109/TVCG.2019.2934784

[191] J. Walker, R. Borgo, and M. W. Jones. Timenotes: A study on effective chart visualization and interaction techniques for time-series data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):549–558, 2016. doi: 10.1109/TVCG.2015.2467751

[192] X.-M. Wang, T.-Y. Zhang, Y.-X. Ma, J. Xia, and W. Chen. A survey of visual analytic pipelines. *Journal of Computer Science and Technology*, 31:787–804, 2016. doi: 10.1007/s11390-016-1663-1

[193] C. Ware. *Information Visualization: Perception for Design*. Elsevier, 2021. doi: 10.1016/C2016-0-02395-1

[194] M. Wertheimer. Laws of organization in perceptual forms. In W. Ellis, ed., *A Source Book of Gestalt Psychology*, pp. 71–88. Routledge and Kegan Paul, London, 1938. doi: 10.1037/11496-005

[195] C. O. Wilke. *Fundamentals of data visualization: a primer on making informative and compelling figures*. O'Reilly Media, 2019.

[196] C. Xiong, C. Stokes, Y.-S. Kim, and S. Franconeri. Seeing what you believe or believing what you see? belief biases correlation estimation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):493–503, 2023. doi: 10.1109/TVCG.2022.3209405

[197] Y. Xu and R. Goodacre. On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of analysis and testing*, 2(3):249–262, 2018. doi: 10.1007/s41664-018-0068-2

[198] F. Yang, L. T. Harrison, R. A. Rensink, S. L. Franconeri, and R. Chang. Correlation judgment and visualization features: A comparative study. *IEEE Transactions on Visualization and Computer Graphics*, 25(3):1474–1488, 2019. doi: 10.1109/TVCG.2018.2810918

[199] L. Yuan, S. Haroz, and S. Franconeri. Perceptual proxies for extracting averages in data visualizations. *Psychonomic Bulletin & Review*, 26(2):669–676, 2019. doi: 10.3758/s13423-018-1525-7

[200] Y. Zhang, M. Fjeld, A. Said, and M. Fratarcangeli. Task-based colormap design supporting visual comprehension in process tomography. In *EuroVis 2020 - Short Papers*. Eurographics, 2020. doi: 10.2312/evs.20201049

[201] Y. Zhang, Y. Sun, L. Padilla, S. Barua, E. Bertini, and A. G. Parker. Mapping the landscape of covid-19 crisis visualizations. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, article no. 608, 23 pages. ACM, New York, 2021. doi: 10.1145/3411764.3445381

[202] H. Zhao, G. W. Bryant, W. Griffin, J. E. Terrill, and J. Chen. Validation of splitvectors encoding for quantitative visualization of large-magnitude-range vector fields. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1691–1705, 2017. doi: 10.1109/TVCG.2016.2539949

[203] A. Zimek, E. Schubert, and H.-P. Kriegel. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 5(5):363–387, 2012. doi: 10.1002/sam.11161

[204] J. Zwinkels. Light, electromagnetic spectrum. In *Encyclopedia of Color Science and Technology*, pp. 1–8. Springer, 2014. doi: 10.1007/978-3-642-27851-8_204-1