Evaluating the effectiveness of uncertainty visualizations

A user-centered approach

Esther Kox

4136977

Master thesis, 27.5 ECTS

Supervisor UU: Ignace Hooge, i.hooge@uu.nl

Supervisor TNO: Ward Venrooij, ward.venrooij@tno.nl

Second reviewer UU: Jeroen Benjamins, j.s.benjamins@uu.nl

Applied Cognitive Psychology

University of Utrecht - 2017-2018

28-06-2018

Abstract

The aim of this research is to compare the effectiveness of several visual representations of statistical data uncertainty and to explore the influence of individual differences among users. A large body of work shows that providing uncertainty information is beneficial for decision-making. However, these advantages of showing uncertainty critically depend on how it is communicated. This large online user study (n=245) identifies how the quality of probability estimates compares across six visualizations and across users. Participants were presented six visualization types that each encode a probability distribution that represents a possible range of arrival times, predicted by a car navigation system. They were asked to report best estimate and two kinds of probability estimates (*later than* and *range)*. Probability estimate accuracy and precision were compared across visualizations, question type and user types based on ten characteristics, among which both cognitive measures and personality traits. An ANOVA showed a main effect of visualization type, where discrete plots with few outcomes result in the most accurate and precise probability estimates and prove to perform best across all user classifications. Although task performance differs across users with different levels of cognitive abilities and personality traits, it can be concluded that visualization type has a much greater impact on performance than individual differences have. This suggests that, when designing an interface with an aim for high performance, it is more effective to focus on the graphic design of a chart than on personalization. The acquired knowledge contributes to the standardization of including uncertainty measures into information visualizations and to the development of user adaptive visualizations.

1.1 Information visualization

In today's society, digital information and computer interfaces have come to play a crucial role in our lives. Data is being collected for a lot of different purposes at an excessive rate. This ever-growing and sometimes overwhelming amount of digital information can be made more manageable with the aid of information visualization. Information- or data visualization (often abbreviated as InfoVis or DataVis) is the use of images to represent data (Few & Edge, 2007). The terms 'information' and 'data' will be used interchangeably, just as the Cambridge Dictionary defines data as a synonym of information: facts or numbers that are collected and examined to support decision-making. Data visualization is used to amplify cognition by the depiction of data, or to aid the exploration of abstract data and the discovery of new insights and knowledge (Chen, 2017). Human vision a powerful tool for data analysis and interpretation, as it is highly selective regarding different sizes, shapes, colors, and spatial positions. In addition, vision enhances memory and cognitive capacity, both of which play a critical role in the way people process information (Chen, 2017). The visualization of information can help people carry out tasks more effectively, as it provides an external representation that replaces cognition with perception (Munzner, 2015). By conserving cognitive resources, it enables people to solve problems that would be hard, if not impossible, to solve if the data was expressed in other forms, like reports and spreadsheets (Lohse, 1997).

As the amount of data grew in the last decades, the advantages of visualization became more clear and InfoVis became a flourishing topic within the field of Human-Computer Interaction (HCI). Despite the increasing interest in how to improve the effectiveness of visualizations, InfoVis neglected HCI's acknowledged process of user analysis and traditionally followed a one-size-fits-all principle. However, in recent years researchers have come to understand that user differences such as cognitive abilities, personality and experience have a significant impact on the effectiveness of visualizations. In the context of user-centered design of information systems, user differences can be thought of as any differences in the resources that users bring to the table during information tasks (Allen, 2000). In the current study, visualization effectiveness is defined as successful in enabling the quick extraction of accurate information (Kennedy, Hill, Allen & Kirk, 2016). For a more detailed description of the concept 'effectiveness', see Appendix A.

Visualization is a powerful tool as it converts plain data into a graphical display that presents large amounts of data in a small amount of space, while expresses the information in a more intuitive, memorable manner (Chen, 2017; Bonneau et al., 2014). But as the amount of data and its complexity grows, it gets harder to effectively and accurately convey the information through a visualization. Data inevitably comes with some degree of uncertainty. Data uncertainty is a broad term that includes various concepts to characterize data. It applies to measurements and observations, as well as predictions and it can involve related concepts like error, accuracy, precision, validity, quality, variability, noise, completeness, confidence, and reliability (Pang, 2001). It can arise in every phase of data analysis, from data acquisition to data visualization (Bonneau et al., 2014). Often, subsets of data are represented by center measures like the mean or median, but the uncertainty associated with such measures, like confidence intervals, variability and model biases, can be as important as the difference between them (Correll & Gleicher, 2014). If visualization is used as a means to assist the exploration of data or to communicate information to others as a base for decision-making, measurements of uncertainty must be included (Griethe & Schumann, 2006). However, due to the lack of solid visualization techniques, uncertainty visualization often remains an unsolved problem (Bonneau et al., 2014).

Every day, people with all kinds of abilities and backgrounds need to make decisions based on digital data presented by interfaces. It is unavoidable that people make mistakes, but optimizing data visualization techniques and human interface interaction can minimize the number of human errors due to bad design. This can be achieved by on the one hand thorough user analysis and user-centred design, and on the other hand by displaying all facets of the data, including its uncertainty, in order to prevent misinterpretation. The current study pays attention to both components, as it explores the effectiveness of six uncertainty visualization techniques while evaluating the potential influence of a wide range of user characteristics.

1.2 Data uncertainty

Definition, sources and examples

Considering the broadness of the term and the many fields the concept is applicable to, it is not surprising that there is no consensus on the precise definition for (data) uncertainty (Pang, 2001). The implications of the multifaceted concept will be illustrated by the following. The sources of data uncertainty can be divided in three broad classes: uncertainty measures generated by models or simulations, uncertainty observed in sampled data, and uncertainty introduced in the phases of data processing or visualization (Bonneau et al., 2014). It all starts with structural uncertainty, resulting from the given that no model can fully grasp or copy the natural world, as it is too complex and abstract (Greis, Schuff, Kleiner, Henze & Schmidt, 2017). The resulting omnipresence of data uncertainty makes that everyone has to deal with data uncertainty in everyday life, without always being aware of it.

People make decisions based on the bus schedule, the weather forecast and the navigation system in their car, even though such information are estimates and will always be partly uncertain due to flawed prediction models, incomplete knowledge, and data noise. In such cases, a measurement- or prediction error could notify a user about the amount of uncertainty associated with the situation.

Likewise, people often attach great value to election polls, while these predictions are based on sampled data: a set of observations assumed to be representative, which is then generalized to the entire population. And as reality teaches, the results and statements that follow from these generalizations must be accepted with caution, since the true responses of the majority of the population will remain unknown, or at least uncertain, until the actual election. Providing a confidence interval would make such polls more trustworthy. Another form of sampled data is the simplification of large or complex datasets. Although all data is known, and thence 'certain', reducing a large dataset to center measures like an average conveys a false simplicity and hides the actual complexity and depth of the data. Including the standard deviation associated with the mean reduces this problem, as it provides information about data variability.

Finally, it is important to understand how data uncertainty spreads or even multiplies during the process of data visualization and how visualization influences the perception and interpretation of uncertainty (Pang, Wittenbrink & Lodha, 1997, Bonneau et al., 2014; Greis et al., 2017). To gain such insight, there must be understood how positive and negative consequences of showing uncertainty weigh up, how perception and cognition influence the interpretation of uncertainty visualization, and what the impact is of differences in audience abilities and backgrounds (Bonneau et al., 2014).

How do positive and negative consequences of showing uncertainty weigh up?

Clearly, uncertainty is an integral part of data. However, uncertainty measures are generally treated as additional variables of multivariate data, instead of presented together with the underlying data (Pang, 2001). It is often omitted in communication to the general public, as it makes it more complex to show the value in a visualization (Hullman, 2016) and for fear it will be misunderstood and misused (Joslyn & LeClerc, 2012). Here, the general public refers to viewers that do not have a deep statistical background or other experience with the concept data uncertainty.

As society's dependence on data increases, the importance of information visualizations that are truthful and complete, while still accessible to the general public, grows. Fortunately, many studies that have been conducted in the past decades scientifically endorsed the relevance of such visualizations and contributed to their development. Empirical research shows that people prefer information that expresses uncertainty (Morss, Demuth & Lazo, 2008) and that the representation of uncertainty information even enables users to make better decisions and increases their trust in the information

(Roulston, Bolton, Kleit & Sears-Collins, 2006; Joslyn & LeCrerc, 2012; Joslyn & LeClerc, 2013; Kay, 2016). Displaying uncertainty can reduce anxiety, as it avoids false precision in single points estimates (Kay et al., 2016) and it benefits decisions by enabling people to anticipate on both the range of possible outcomes and the amount of uncertainty associated with the situation. However, these advantages of showing uncertainty critically depend on how it is communicated (Joslyn & LeClerc, 2013).

How do perception and cognition influence the interpretation of uncertainty visualization?

Although the importance of integrating uncertainty into a data visualization is increasingly recognized, it remains a challenge to get it right. Data and its associated uncertainty, like sample mean and error, are traditionally represented by a bar chart with error bars, especially in the scientific world. However recently, error bars received a lot of critique due to their severe shortcomings, among which several perceptual biases (Correll & Gleicher, 2014). The drawbacks associated with error bars and the evaluation of alternative encodings for uncertain information that are proposed in literature are discussed in Appendix B. The perceptual properties like color, size, and location (in InfoVis literature often referred to as *retinal variables*) that influence the interpretation and effectiveness of an encoding will there be compared by using Mackinlay's (1986) ranking of effective visual encodings. Also, a distinction is made between intrinsic and extrinsic annotations of uncertainty. Error bars are an example of an *extrinsic* annotation of uncertainty, meaning that the underlying data and the corresponding uncertainty are not integrated into the same encoding. According to the heuristics that are employed in making judgements under uncertainty, the separation increases the risk of the extrinsic uncertainty representation being perceived as peripheral, and of later being discounted when making judgements (Tversky & Kahneman, 1974). An *intrinsic* annotation of uncertainty on the other hand integrates uncertainty values into the underlying data, which avoids ambiguity and simplifies interpretation (Kay, 2016). Kay (2016) argues that to encourage viewers not to undervalue probability information, uncertainty should be intrinsic to the representation of the underlying data. But visualization effectiveness does not only depend on the visualization itself, it depends on the cognitive goal of the user as well; on the information that a user is trying to extract from the visualization (Ibrekk & Morgan, 1987).

What is the impact of differences in audience abilities and backgrounds?

Essential to creating an effective uncertainty visualization is understanding that goal; understanding why uncertainty needs to be visualized and in what way the uncertainty visualization needs to help the user. Following this user-centered approach might mean that the same data should be visualized differently for different audiences (Lapinski, 2009). Before getting deeper into the impact of user differences on visualization effectiveness and the importance of user-centered design, the following section will evaluate different types of uncertainty and different techniques for visually representing it.

1.2.1 Techniques for visualizing statistical uncertainty

Types of data uncertainty

Different types of uncertainty require different types of visualization techniques, as a recent experimental study showed that effectiveness also depends on the *type* of uncertainty represented (Gschwandtner, Bögl, Federico & Miksch, 2016). Olson & Mackinlay (2002) distinguish two types of uncertainty: *statistical uncertainty* and *bounded uncertainty*. In case of statistical uncertainty, the probabilities of values within an potentially infinite distribution depend on a statistical model that includes assumptions about the most likely point of estimate. Whereas with bounded uncertainty, not one value has the highest probability. Instead, all values that lie inside a bounded range, defined by precise lower and upper bounds, are equally likely (Olson & Mackinlay, 2002).

Use case based on statistical uncertainty

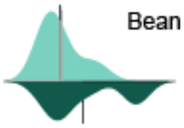
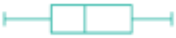
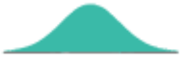
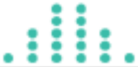
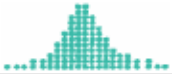
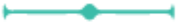
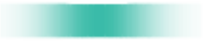
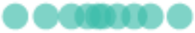
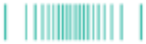
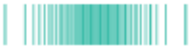
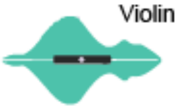
The use case that will be adopted in the current study simulates a car navigation system that recommends a potential route with a corresponding estimated time of arrival. However, providing a fixed point estimate time of arrival without presenting the associated range of uncertainty (prediction error) will often convey a false precision. Providing a probability estimate can support users in understanding that there is a chance that they will arrive earlier or later than the point estimate and can help them assess schedule opportunities (Kay, Kola, Hullman & Munson, 2016). The estimated time of arrival by a navigation system is prone to uncertainty, due to potential obstacles or thanks to potential windfalls down the road. The goal is to find a visualization that effectively communicates this uncertainty. The current study will focus solely on the visualization of statistical uncertainty. Given the fixed departure time, travel distance, prescribed speed limits and the possibility to take these potential influential events into account, statistical assumptions can be made about a best point of estimate and its corresponding uncertainty. The range of possible arrival times is not absolute or bounded; in the worst case scenario, the driver will never arrive at his/her destination. Meeting the requirements for statistical uncertainty mentioned above, the uncertainty associated with navigation arrival time can be considered statistical uncertainty, following a normal distribution. The distribution can slightly shift as a result of traffic lights and can be skewed as a result of events with more heavy consequences, like a traffic jam.

Selecting visual encodings of statistical uncertainty

Table 1 gives an overview of different ways to visually encode uncertain statistical information that were found in the literature. Here, all visualization types are adjusted to a horizontal layout, to resemble a timeline. The criteria are based on the literature and the requirements associated with the chosen use case. Based on literature, an intrinsic annotation of uncertainty is desirable and makes summary statistics superfluous. For this study, visualizations were selected that could effectively encode probability in relatively small spaces. If the visualization would actually be implemented in a car navigation system, the encoding should be able to convey its information on a small screen. Therefore, visual clutter risks must be avoided. Finally, the visualization should be able to effectively and intuitively communicate a probability distribution to the general public, without the need for prior knowledge.

Table 1.

An overview of criteria for the visual encodings of probability that were considered for the study.

	Intristic annotation	Summary statistics	Directly estimate density	Easily assess mode	Low risk of visual clutter	No need for prior knowledge
Bean 	●	●	●	●		●
Box plot 	●	●			●	
Density 	●		●	●	●	●
Dot (20) 	●			●	●	●
Dot (100) 	●		●	●		●
Error bars 		●		●	●	●
Gradient 	●		●		●	●
Strip 	●					●
Stripe (20) 	●				●	●
Stripe (50) 	●		●			●
Violin 	●	●	●	●		

The six visualization techniques that were selected for the current study will be mentioned shortly in the following section. For more detailed descriptions and literary backgrounds of all the visualization techniques that were considered and selected, consult Appendix B.

Error bars

Error bars are frequently used, cap-tipped lines that serve as an external, graphical enhancement to display the uncertainty of the plotted data (Figure 1). They can be applied to scatter plots, dot plots, line graphs, and bar graphs. The lines span 95% of the associated probability distribution, which follow the theoretical idea that values closer to the upper and lower boundaries are less likely than values to the reported point estimate (middle).

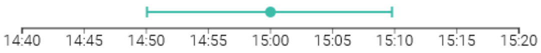


Figure 1. Error bars plot

Density plot

A density plot is a function graph of a probability distribution function and encodes the density as distance from the x-axis (Figure 2) (Kay, 2016). The resulting curve (or *area*) provides a simple summary of the distributions shape and enables quick visual interference about the distribution of the data. To convey the probability density, the density plot relies on the width of an area.

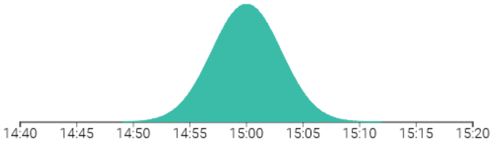


Figure 2. Density plot

Dot plot

A dot plot shows discrete quantiles on a continuous scale using a dot or other symbol (Figure 3) (Wilkinson, 1999). By stacking the dots, the plot shows the distribution of the data, but does not include any graphical description of summaries (Benjamini, 1988). However, the possibility to manipulate the amount of dots used to represent the data, makes summary statistics unnecessary.

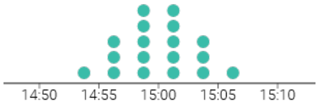


Figure 3. Dot plot (Dot-20)

Stripe plot

The stripe plot is a variation on the strip plot; a one-dimensional scatter plot representing individual observations or probabilities using a dot, or in case of the stripe plot a stripe (Figure 4). Probability density is thus encoded by the density of vertical stripes in a region (Kay, 2016).

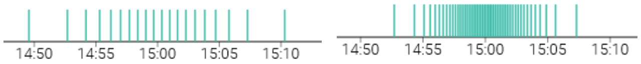


Figure 4. Stripe plots (left: Stripe-20, right: Stripe-50)

Gradient plot

A gradient plot is a shaded horizontal bar glyph of fixed height and width, in which the probability density of the quantity at a point is encoded by opacity (Figure 5). The darker the shade, the higher the probability of the given estimate (Jackson, 2008).

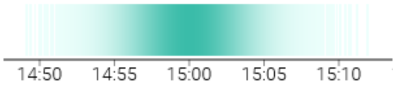


Figure 5. Gradient plot

1.3 User characteristics

Despite the large body of work on uncertainty visualization, knowing when to use which visualization remains an unresolved issue as most studies are very context-specific (Greis, Joshi, Singer, Schmidt & Machulla, 2018). The current study contributes to the body of knowledge by providing insight into the influence of the factor which is present is every context; the influence of the viewer on visualization effectiveness. Design serves as the communication between object and user. User-centered design, a design process that focuses on the users' needs and requirements (Norman, 1988), is therefore a relevant and important concept in developing information visualizations with the purpose of clear communication. In the past decades, researchers have come to understand that the effectiveness of information visualizations not only depends on the visualization itself but is influenced by the characteristics of a user as well. The idea that effectiveness can be boosted through personalization motivated researchers to explore which user features are worth adapting to. The individual differences that might impact the interaction between user and visualization include cognitive abilities, personality, and chart expertise (Toker, Steichen, Gingerich, Conati, & Carenini, 2014) and will be discussed in the next section. Most of these findings, however, are based on studies that use visualizations without any notion of uncertainty. At the same time, earlier work that focuses on the effectiveness of uncertainty visualizations often ignores user variation. The few studies that include both uncertainty in their stimuli and the impact of user characteristics in evaluating the effectiveness of data visualizations address the user features numeracy, level of education and chart expertise. As the literature will be discussed in the following section, it will become clear that the current study is a valuable addition to prior work.

Table 2.

An overview of the user characteristics that will be assessed in the current study.

User characteristic	**Definition**
Perceptual speed	A measure of speed when carrying out all sorts of simple tasks involving visual perception (Conati & Maclaren, 2008).
Visual working memory	The part of the working memory responsible for temporary storage and manipulation of visual and spatial content (Logie, 1995).
Verbal working memory	The part of the working memory responsible for temporary storage and manipulation of verbal information (Baddeley, 1986)
Numeracy	The ability to process, communicate and interpret numerical information in a range of contexts and to solve a variety of problems. (Askew, Rhodes, Brown, William, Johnson,1997)
Conscientiousness	A personality trait defined as the propensity to follow socially prescribed norms for impulse control, to be goal directed, to plan, and to be able to delay gratification. (Roberts, Jackson, Fayard, Edmonds & Meints, 2009).
Extraversion	A personality trait that represents the degree to which a person is open-minded, action-oriented and seeks the society of others. (Green & Fisher, 2010)

Neuroticism	A personality trait distinguished by negativity and a propensity to be moody. (Green & Fisher, 2010)
Locus of Control (LOC)	The degree to which individuals attribute life events and outcomes as either a result of their own behavior, or of forces that are external to themselves. (Green & Fisher, 2010)
Need for Cognition	An individual's tendency to engage in and enjoy effortful cognitive activities. (Wu, Parker & De Jong, 2014)
Self-esteem	The extent to which one prizes, values, approves, or likes oneself (Blascovich & Tomaka, 1991)
Chart expertise	Expertise refers to the level of experience a person has in a craft and the use of a given set of technologies (Green, Jeong & Fisher, 2010) Chart expertise is expertise in a specific visualization or chart type.

Cognitive measures

The comprehension of information visualizations involves information processing and reasoning. Cognitive abilities shape the way we think and how we carry out all sorts of tasks, ranging from simple to complex. Hence, task performance is likely to depend on it.

An important cognitive measure when it comes down to visual perception is perceptual speed. Earlier work shows that it is possible to predict which visualization will be most effective for a person, based on their level of perceptual speed (Allen, 2000; Conati & Maclaren, 2008). Other literature shows that users with high perceptual speed are faster in completing visualization tasks than users with low perceptual speed (Toker et al., 2012; Conati, Carenini, Hoque, Steichen & Toker, 2014; Carenini et al., 2014).

The visual working memory is a part of the perceptual and cognitive processing system where external visual information that enters is briefly stored (Patterson et al., 2014). Hence, a user's visual working memory capacity is involved in the ability to process visualizations. In visualization tasks, visual working memory positively correlates with task accuracy and negatively correlates with completion time

(Velez et al., 2005; Toker & Conati, 2014). Other studies show that users with different levels of visual working memory benefit from different forms of spatial layout in terms of task performance (Conati et al., 2014) and that it affects visualization preference (Toker et al., 2012).

Information visualization often include both graphs and text, intended to complement each other. Therefore, the influence of verbal working memory capacity is studied as well. Eye tracking studies show that users with high verbal working memory consult the textual areas of visualizations less frequent and more quickly than users with low verbal working, who need more time processing text in the task questions as well (Steichen et al., 2013; Toker & Conati, 2014)

An interesting overarching result on the three cognitive measures perceptual speed, visual working memory, and verbal working memory shows that these measures have no significant impact on task performance during simple tasks, while for complex tasks, participant with high scores on these measures performed significantly better. This suggests that user performance depends on cognitive abilities more heavily as task complexity increases (Carenini et al., 2014).

A recent study that focused on the perception of visual uncertainty representations showed that a person's numeracy affects the way an uncertainty range is interpreted (Tak, Toet & van Erp, 2014). Their results show an interaction between the degree of uncertainty and numeracy on perceived probability, where participants with relatively high numeracy have a slightly more extreme interpretation than those with lower numeracy.

Personality

Besides cognitive abilities, every user has its own unique personality that he or she brings to the interface. Personality factors are inherent individual differences, and some have shown to influence information visualization effectiveness

The Big Five factors of personality have been broadly accepted within psychology literature for decades. Green & Fisher (2010) report that users with higher levels of extraversion were faster in finding target information than users with lower levels of this personality factor. The same study shows that users with a lower level of extraversion reported more insights, a measure of how many new things were learned during the task. Similar results were found for the Big-Five personality trait neuroticism. As for extraversion; higher levels of neuroticism resulted in faster task completion times and lower levels of neuroticism led to the report of more insights (Green & Fisher, 2010). In a preliminary study, extraversion was positively correlated with task performance in a simple visualization task (Venrooij, 2018). Remarkable is that researchers do not often motivate why some user traits are included in studies and some excluded for that matter, while especially personality traits sometimes seem far-fetched in perceptual and cognitive research. The Big Five factor of personality Conscientiousness has not yet been included in InfoVis research, but could possibly affect the way in which people read charts and estimate probabilities, since is has to do with being exact and dutiful. Therefor, it is included in the current study.

In InfoVis literature, the personality trait Locus of Control (LOC) is often considered. People with a more internal LOC hold an inherent belief that events and outcomes are under a person's control, and thus, success or failure depends largely on personal behavior and attitudes. The inherent belief that events and outcomes are influenced by external factors is associated with a more external LOC (Green & Fisher, 2010). Several visualization studies showed significant effects of LOC on task performance, while an explanation for the effects remained elusive (Green & Fisher, 2010). Ziemkiewicz et al. (2011) argued that it seems unreasonable that a personality trait without any known connection to visual or spatial ability should have any constant impact over such a complex relationship. The researchers showed that the known effects of LOC can still be found when restricting visualization differences to layout factors. This indicates that LOC relates to the way a user approaches the external representation rather than to how a user visually processes the visualization. They suggest that LOC influences the way people use various visualization types by affecting a user's willingness to adapt to new externalization of information. Likewise, LOC might influence task performance in visualization tasks including uncertainty. Since uncertainty is often omitted in information visualizations, people are rarely exposed to it. The inclusion of uncertainty in the current study might therefore demand the user to adapt to a new externalization of information. Hence, a user's willingness to adapt to new visual encodings, associated with a user's LOC, might influence task performance in the current study.

The personality trait Need for Cognition refers to an individual's tendency to engage in effortful cognitive activities. Hence, it might influence task performance in the cognitive graph task in the current study. Conati & Maclaren (2008) found that Need for Cognition is a positive predictor of user accuracy in a visual sorting task. Lastly, Self-esteem is a measure of self-evaluation of one's social identity, worth and value (Blascovich & Tomaka, 1991) and is included as a standard variable in social construct studies.

Chart expertise

Besides cognitive abilities and personality, individual differences also arise from different levels of experience. Earlier work shows that higher expertise with a specific visualization type or task is positively correlated with visualization task accuracy (Lewandowsky & Spence, 1989). Expertise has also shown to be a predictor for visualization preference; the higher the expertise with radar graphs, the higher the preference (Toker et al., 2012). Moreover, users with high expertise are faster compared to users with low and average expertise when completing a complex sort task (Conati et al., 2014).

User variation in uncertainty studies

Except for the study on the effects of numeracy, all of the results mentioned above are based on visualization tasks without data uncertainty. Only a few studies look at the potential effects of user variation when evaluating visualizations with uncertainty. In contrast to the findings mentioned above, studies that evaluate the effect of expertise on the perception of uncertainty in various visualizations find no significant difference in performance when comparing novice to experts (Evans, 1997; Blenkinsop, Fisher, Bastin, & Wood, 2000; Aerts, Clarke, & Keuper, 2003). Ibrekk & Morgan (1987) found that an active knowledge of statistics, associated with experience in uncertainty visualizations, mitigated the most obvious misinterpretations when using an uncertainty visualization, but did not necessarily result in a higher task performance.

Other studies especially focus on whether or not providing uncertainty information can benefit the decision-making of non-experts at all, without examining the effects of level of expertise on decision-making (Roulston, Bolton, Kleit & Sears-Collins, 2006; Joslyn & LeCrerc, 2012; Joslyn & LeClerc, 2013). Similarly, a health care study from 2001 studied how receptive patients are for uncertain, probabilistic information and found that the education level of a patient may influence both the understanding of, and the receptivity for, uncertainty (Schapira, Nattinger & McHorney). A group of women was presented visual depictions of a points estimate with a confidence interval associated with the risk reduction of breast cancer mortality. The more educated women (with at least a 4-year college degree) were accepting of ambiguity, and most of them felt that the CI should be presented, whereas the less educated women perceived the information as less trustworthy and generally desired the information to be conveyed in a simpler format.

Thanks to the growing awareness that uncertainty is relevant and can benefit decision making, efforts are being made to make uncertainty accessible to the wider public. As an illustration, a group of researchers developed a toolkit meant to make trustworthy national metrics data available to policy makers, journalists, the well informed public, and ultimately students at every educational level (Daken, Dogruel, Grimes, Lam & Lotze, 2008). In order to do this right, the developers emphasize that user evaluation is crucial. However, the influence of user characteristics is one of the critical areas missing from the literature reviewed on visualizing uncertainty, they argue. Gherson (1998) states in his piece

‘Visualization of an Imperfect World’ that user variation is one the things that makes visualization challenging. Since no two users are alike, methods must be developed that allow visualization to be personalized.

The current study contributes to the search towards effective visualizations of uncertainty and combines this with elaborating on the findings that visualization effectiveness depends on user characteristics. This study evaluates task performance on a task that requires not only low-level visual processing but also high-level cognitive treatment of the visual information and measures the accuracy of probability estimates with six visualizations of statistical uncertainty and considers user variation in a broad sense, including the influence of perceptual speed, visual working memory, verbal working memory, numeracy, extraversion, neuroticism, locus of control, chart familiarity and education. The current study uses chart *familiarity* instead of chart *expertise*, since the presented visualizations of uncertainty are rather new and unknown. Chart familiarity will be evaluated as a subjective measure, together with visual appeal and ease-of-use, since it is self-reported and not objectively measured. By comparing six visualizations that represents statistical data uncertainty, the goal of this study is to identify the one(s) that are best suited for conveying probability information to the general public, while also exploring if uncertainty visualization effectiveness might differ across user types, based on several cognitive- and personality characteristics. This will be investigated through an online user study. In which participants will be shown visual representations of an estimated time of arrival within a range of associated uncertainty, which is, according to the use-case, provided by a car navigation system. For each case, there are three task types: judge the best estimate time of arrival, estimate the chance that arrival time will be later than a given point in time, and estimate the chance that arrival time will fall in between a certain time range.

1.4 Research questions & Hypothesis

Research questions

RQ1: How do accuracy and precision in probability estimates compare across six visualizations that represent statistical data uncertainty ?

RQ2: Is visualization effectiveness influenced by user characteristics? Are some visualizations better suited for specific user types?

RQ3: How is task performance (measured in percentage correct) related to subjective measures like chart familiarity and ratings of visual appeal and ease-of-use?

Hypothesis

Uncertainty visualization effectiveness depends on task type. As Ibrekk & Morgan (1987) found, performance depends upon the information that a subject is trying to extract from a visualization. Visualizations that explicitly contain the information that people need result in the best performance.

Probability estimates that are based on the visualization Errorbars are significantly less accurate than with the other visualizations, because the finer details about the probability distribution are hidden in error bars (Cairo, 2016).

Probability estimates that are based on the visualizations with discrete outcome plots with few enough outcomes to benefit from subitizing, like the Dot-20 and Stripe-20, are significantly more accurate than the other visualizations (Kay et al., 2016).

Users with high perceptual speed perform better across all visualizations (Carenini et al., 2014)

Users with high scores on the cognitive measures PC, visual WM and verbal WM perform significantly better on Q2 and Q3 than users with lower scores. There is no difference in performance for Q1. Carenini et al. (2014) suggests that user performance depends on cognitive abilities more heavily as task complexity increases.

Users with higher scores on numeracy exhibit more pronounced estimating behavior than users with lower scores on numeracy, i.e. overstating high probabilities, while understating low probabilities. (Tak, Toet & van Erp, 2015).

Users with a more external locus of control perform better with these relatively unknown visualizations including uncertainty. According to earlier work, they will have a higher willingness to adapt to new externalization of information (Ziemkiewicz et al. 2011).

Users with high scores on extraversion perform better than users with low scores on extraversion (Green & Fisher, 2010; Venrooij, 2018) and users with high scores on neuroticism perform better than users with low scores on neuroticism (Green & Fisher, 2010).

Familiarity with a visualization is not related to task performance (Evans, 1997; Blenkinsop, Fisher, Bastin, & Wood, 2000; Aerts, Clarke, & Keuper, 2003).

Users with higher familiarity with a certain visualization rate that specific visualization higher in visual appeal than users with low familiarity. In other words, familiarity with a visualization type positively correlates with that visualizations rating on visual appeal. This hypothesis is formulated to test the suggestion of Kay et al. (2016) that Dot-20 might be rated lower in visual appeal than the density plot due its relative unfamiliarity, while it showed to be ~1.15 times more precise than the density plot and yielded higher confidence.

2. Method

To answer the research questions, an online user study was conducted. In the following, the experimental design, task, stimuli, and the procedure for the experiment is presented.

2.1 Experimental design, stimuli and task

Design

A 10 x 6 x 3 within-subject design was used with 3 independent variables: visualization type (with the six levels Density, Dot-20, Gradient, Stripe-50, Stripe-20 and Errorbars), task/question type (with the three levels best estimate, later than probability and range probability), and user characteristics (with the ten levels perceptual speed, visual working memory, verbal working memory, numeracy, conscientiousness, extraversion, neuroticism, need for cognition, locus of control), and the dependent variable: percentage correct in probability estimates.

Stimuli: Visualizations and questions

The six types of visual representation of a probability distribution that are evaluated in the current study are depicted in Figure 6. The set consists of three continuous visualization types (Density, Gradient and Errorbars) and three discrete ones (Dot-20, Stripe-20 and Stripe-50). The horizontal error bar is inspired by the study of Ibrekk & Morgan (1987). Three encodings are adopted from the study of Kay et al. (2016) that aimed to identify effective visual encodings to convey the uncertainty associated with bus arrival times to users. Since the current study shares this goal and considers a similar traffic use case, the following visual encodings are adopted: the density plot, the dot plot (20) and the stripe plot (50). The gradient plot in timeline format is adopted from the study of Gschwandtner et al. (2016). Finally, the Stripe-20 is introduced. All visualization are made with D3.js.

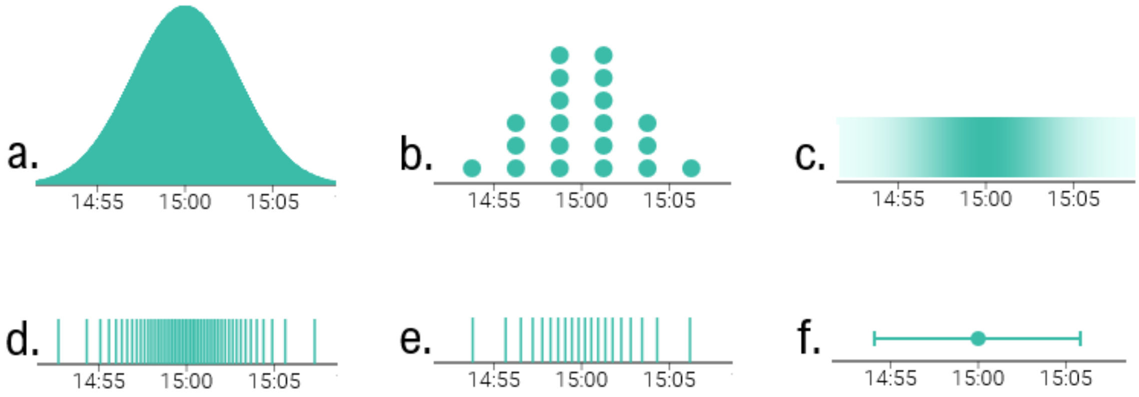
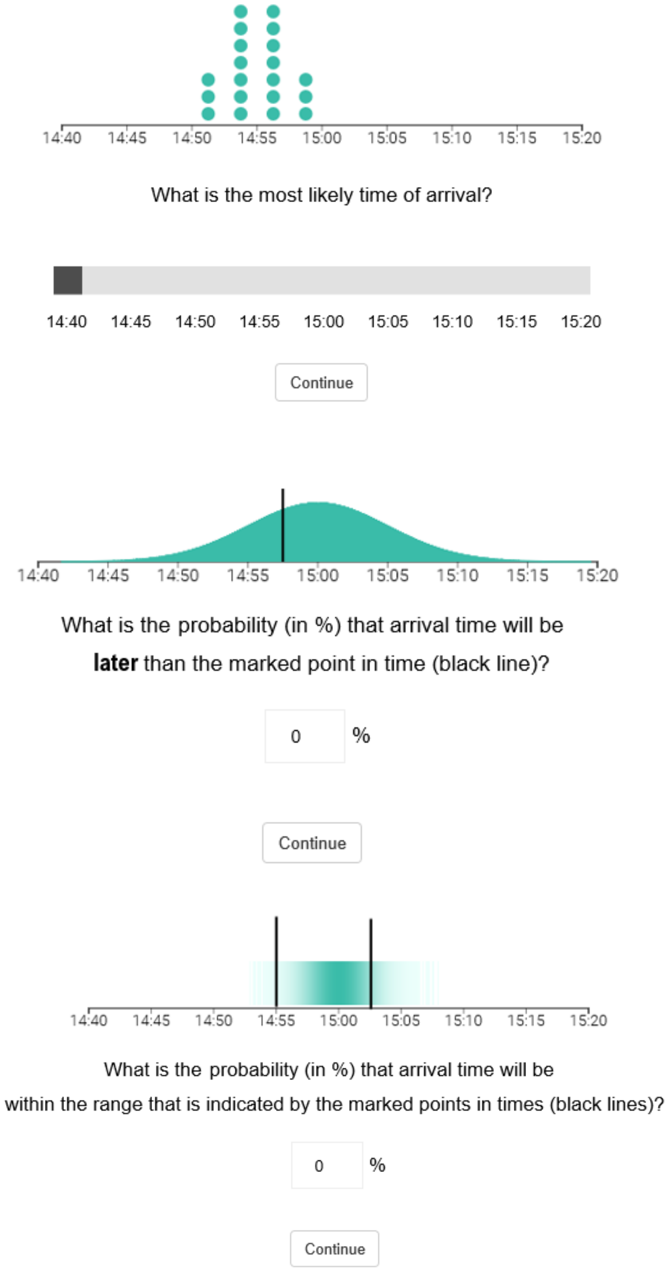


Figure 6. The six types of visualizations selected for evaluation (mean 0, std.dev. 1). 1a. Density; 1b. Dot-20; 1c. Gradient; 1d. Stripe-50; 1e. Stripe-20, 1f. Errorbars

To assess people's ability to judge probability from the visualizations, a similar approach to that of the study of Ibrekk and Morgan (1987) was used. They showed various representations of uncertainty for weather forecasts and participants were asked to report the best estimate and two kinds of probabilities (e.g., snowfall >2 inches, or between 2 and 12 inches). Hence, the current task consisted of the following three questions:

- Q1: What is the most likely time of arrival?
- Q2: What is the probability (in %) that arrival time will be later than the marked point in time (black line)?
- Q3: What is the probability (in %) that arrival time will fall within the range indicated by the marked points in time (black lines)?

Participants were asked to answer the first question by selecting a time on a visual scale slider and the second question by entering a percentage (Figure 7).



What is the most likely time of arrival?

What is the probability (in %) that arrival time will be **later** than the marked point in time (black line)?

0 %

What is the probability (in %) that arrival time will be within the range that is indicated by the marked points in times (black lines)?

0 %

Figure 7. Example screenshots of the experiment (From the top down: Q1 with Dot-20; Q2 with Density plot; Q3 with Gradient)

For each question, there were 48 items; eight variations of a normal standard distribution, displayed by six visualizations. For the sake of divers responses, the variations are adapted to the nature of the question (i.e. Q1 depends on the mean, Q2 and Q3 depend on the distribution form determined by the standard deviation). For question 1 there were 8 combinations of mean -1, -0.5, 0, 1 with standard deviations of 0.5 and 1. For questions 2 and 3 there were 8 combinations of mean 0 and 1 with standard deviations of 0.4, 0.6, 0.8 and 1.

User characteristics

The tests that were used to administer all user characteristics can be found in Table 3.

Table 3.

8 psychometric measures were administered

Psychometric measure	**Test**
Perceptual speed	The Number Comparison test (P-3) to assess perceptual speed from Ekstrom (1976) was digitized for the current study
Visual working memory	Fukuda & Vogel's colored squares test (Fukuda & Vogel, 2009)
Verbal working memory	Operation-word span test (OSPAN) (Turner, 1989)
Numeracy	An abbreviated Numeracy scale (Weller et al., 2013)
Extraversion	the IPIP 10-item Big Five Extraversion Scale (Donnellan et al. 2006)
Neuroticism	the IPIP 10-item Big Five Neuroticism Scale (Donnellan et al. 2006)
Conscientiousness	the IPIP 10-item Big Five Conscientiousness Scale (Donnellan et al. 2006)
Locus of Control	the IPIP 10-item Locus of Control scale: Internality (Levenson, 1981)
Need for Cognition	The 10-item Need for Cognition Scale (Cacioppo & Petty, 1982)
Self-esteem	The 10-item Rosenberg self-esteem scale (RSE) (Rosenberg, 1965)

Information about a participants highest level of education completed is included in their Prolific account (seven categories: 1 = primary education/no education; 2 = lower vocational education; 3 = lower secondary education; 4 = higher secondary education; 5 = BSc; 6 = MSc; 7 = PhD).

Task

The online study was divided into three components: (1) administration of the user characteristics, (2) the main experiment and (3) a post-questionnaire. Further information about part 1 can be found in Table 3.

At the start of the main task, a brief explanation about the purpose of the visualizations was given. The instruction was the following:

"*Imagine you are in a car. Your navigation system recommends a route to your destination. Instead of a fixed estimated time of arrival, a visualization is presented. The visualization represent the probability of arriving at a certain point in time, on top of a timeline. During the task, you will see six different visualizations. For each visualization, you are asked several questions. A brief explanation of the six visualization follows."*

Then, the images of Figure 6 and 8 were presented with the following brief instruction on each visualization:

*"In a normal distribution, values closer to the upper and lower boundaries are less likely than values to the reported point estimate. The image below (Figure 8) shows how probability is distributed. In visualization **a**, the cap-tipped lines span 95% of all values and the arrow represents the point of 50%. In visualization **b**, probability is depicted by the height of the curve, as shown in Figure 8. In visualization **c**, probability is represented by the amount of dots. It consists of 20 dots, meaning that every dot represents 5%. Visualization **d** communicates probability by the density of vertical stripes in a region. It consists of 50 stripes, meaning that every stripe represents 2%. Visualization **e** communicates probability by the density of vertical stripes in a region. It consists of 20 stripes, meaning that every stripe represents 5%. Visualization **f** conveys probability by opacity; the darker the shade, the higher the probability of the given estimate."*

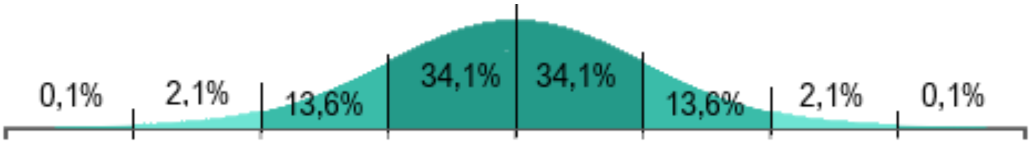


Figure 8. The probability distribution of a normal distribution depicted by a density plot, or bell curve.

Prior to the actual task, participants got 18 practice trials (6 visualization types x 3 questions) that allowed them the opportunity to get familiar with the questions and visualization types, while also stabilizing a potential practice effect. The actual experiment consisted of three question 'rounds'. Within each round, the 48 stimuli per question were presented to the participants in random order, resulting in an 144 item task in total. Each participant answered all questions for all types of visual encodings.

Afterwards in a post-questionnaire, chart familiarity was self-reported by the participants by expressing their agreement with the following statement for each visualization type: "I am familiar with the density plot" on a Likert-scale from 1 to 5. Lastly, participants were asked to rate the ease-of use and visual appeal for each visualization on a Likert-scale from 1 to 5.

2.6 Participants

In total, 245 subjects, participated (101 male, 144 female, age 18-62 years with a mean of 36,2). Participants were recruited via the online academic database Prolific. All participants were native English speakers. Education level frequencies are reported by figure 9 . Initially 323 subjects participated, but 48 subjects were eliminated based on their response patterns (e.g. repetitive answers over longer period of time and unreliably fast reaction times), another 4 subjects scored lower than 10% correct on the graph task and 1 subject was eliminated because he/she took over 4 hours to complete the task. Moreover, two questions that were added to the personality questionnaires to see if people were paying attention ("*are you paying attention? Please answer "Strongly agree" to this question*") were incorrectly answered by 10 people, whom were consequently eliminated. Lastly, 15 subjects were eliminated based on their performance on the cognitive tasks. Their scores were that low that it is thought to be safe to assume that they were not making a serious attempt while performing the tasks. This results in 245 participants.

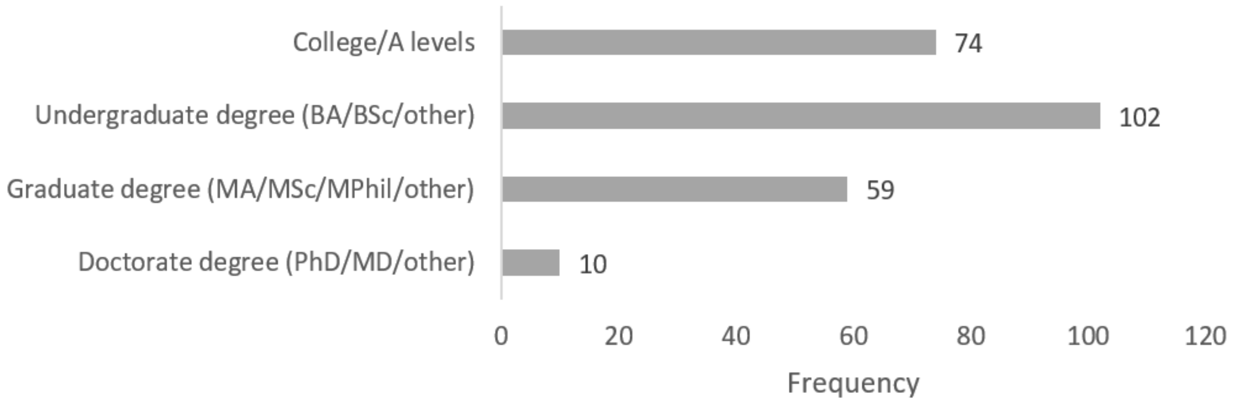


Figure 9. Frequency distribution of subjects their education levels, total n=245.

Payment

Participants performed the study online and were paid a base rate of £6 for their work. They were told that a bonus payment of £1.00 would be given for effortful responses. 245 participants ultimately received this bonus.

2.7 Statistical analysis

For the analysis of the data obtained during the graph task, the response of the participant will be referred to as the *estimated p* and will be compared to the *true p*, which is calculated from the underlying probability distribution of the visualization. For more detailed explanations on the outcome measures, consult Appendix A.

To evaluate the effect of visualization type, a repeated measures ANOVA will be performed to compare the average performance (in percentage correct) for each visual encoding and to see which visual encoding scored best. Bivariate correlation coefficients will be calculated to assess the size and direction of the relationships between task performance in general, task performance per visualization type and user characteristics scores. A regression will be performed to analyze if variation in task performance can be attributed to variation in user characteristics scores.

3. Results

In analyzing the results of the graph task, *true p* is subtracted from *estimated p* to calculate *error size*. Error size is a measure of the *accuracy* of the responses. A response is considered 'correct' when error size is equal to or less than one; *correctness* is a binary scale. Then, the error size distribution has been used to evaluate the *precision* per visualization type. A more detailed overview of these concepts can be found in Appendix A.

Task type

As shown in Figure 10, question 1 (*What is the most likely time of arrival?*) shows a different pattern than questions 2 and 3 (*What is the probability that arrival time will - be later than the marked point in time? / ~ fall within the marked points in time?*). Probability estimates elicit a different response pattern than judging the best estimate. Looking at Table 4, average reaction time seems to be determined more by question type than by visualization type. As hypothesized, performance depends task type; upon the information that a subject is trying to extract from a visualization.

For further analysis of the effect of visualization type and the influence of user characteristics, question 1 will be left out of consideration since there is too little variation as almost every visualization yielded responses near 100% correct. The result patterns of question 2 and 3 have proved to be robust through all pilots and will be considered for further analysis.

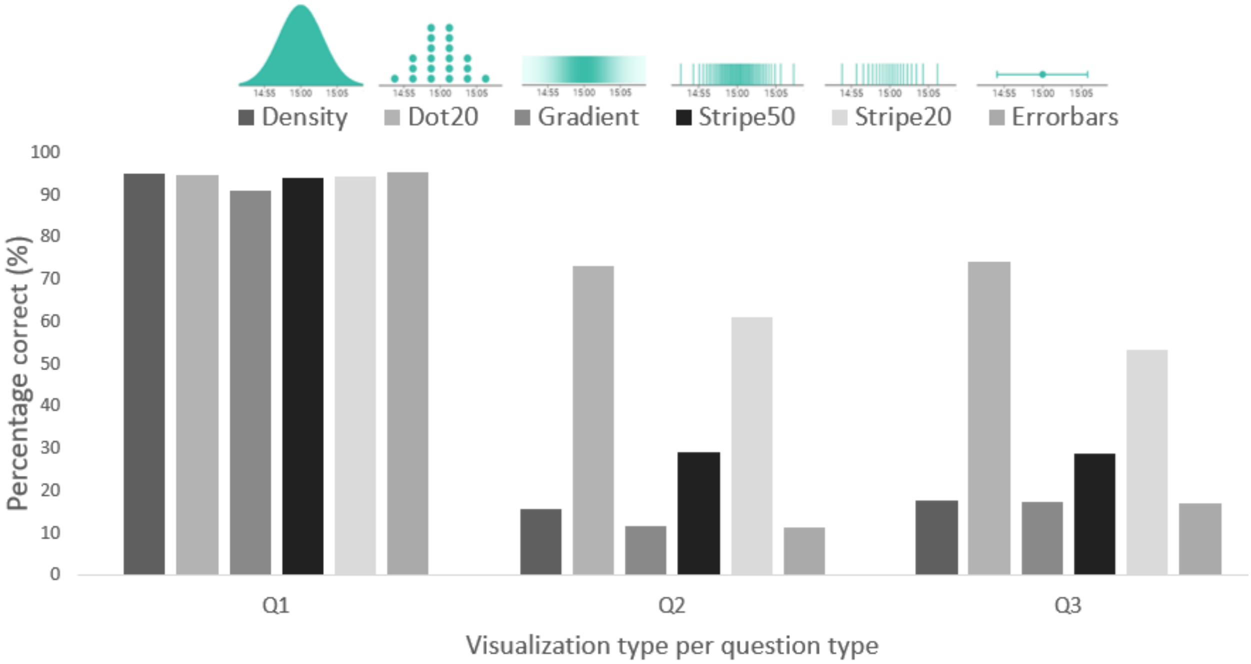
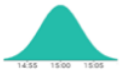
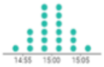
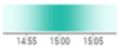
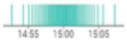
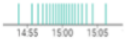


Figure 10. Percentages correct on the graph task per visualization per question type (n=245). For left to right Q1, Q2, and Q3.

Table 4.

Average reaction times per visualization per question in seconds.

Visualization type		Q1 RT (s)	Q2 RT (s)	Q3 RT (s)
	Density	5,1	12,1	8,8
	Dot-20	5,4	10,7	9,1
	Gradient	5,3	10,8	9,1
	Stripe-50	5,2	13,5	12,0
	Stripe-20	5,6	10,4	9,2
	Error bars	5,0	11,1	9,0
	Mean	5,3	11.4	9,5

Visualization type

Comparing visualization types based on accuracy

The data of questions 2 and 3 were further analyzed using a repeated measures analysis of variance (ANOVA) with the within-subjects factor visualization type with six levels (Density, Dot-20, Gradient, Stripe-50, Stripe-20, and Errorbars). Statistical significance is reported at the 0.05 level, as well as partial eta squared (ηp²) for effect size, where .01 is a small effect, .09 is a medium effect, and .25 is a large effect (Toker et al., 2012).

Shapiro-Wilk statistics indicated that the assumption of normality was violated (sig < .001). Fmax = 8.132 demonstrates homogeneity of variances. Mauchly's test indicated that the assumption of sphericity had been violated (χ2(2) = 501.925, p< .001), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε= 0.481).

The ANOVA results shows that there is a main effect of visualization type on percentage correct, F(2.41, 586,96) = 492,04, p < .001, ηp2= .668. A Bonferroni corrected post hoc test further revealed that the Dot-20 results in the highest percentage correct with a significantly higher average than all other visualizations (p < .001) (Figure 11). The descriptives per visualization can be found in Table 5. Following up is the Stripe-20, which significantly differs from all visualizations, including the dotplot (p < .001). Stripe-50 also differs significantly from all other visualization (p < .001). The bottom three visualizations; Density, Gradient and Errorbars do not differ significantly from one another.

Table 5.

Descriptives of the percentages correct per visualization type.

	M	SD
Dot-20	74.3	31.7
Stripe-20	57.8	30.3
Stripe-50	29.5	19.8
Density	16.9	12.7
Gradient	14.6	11.1
Errorbars	14.1	11.6

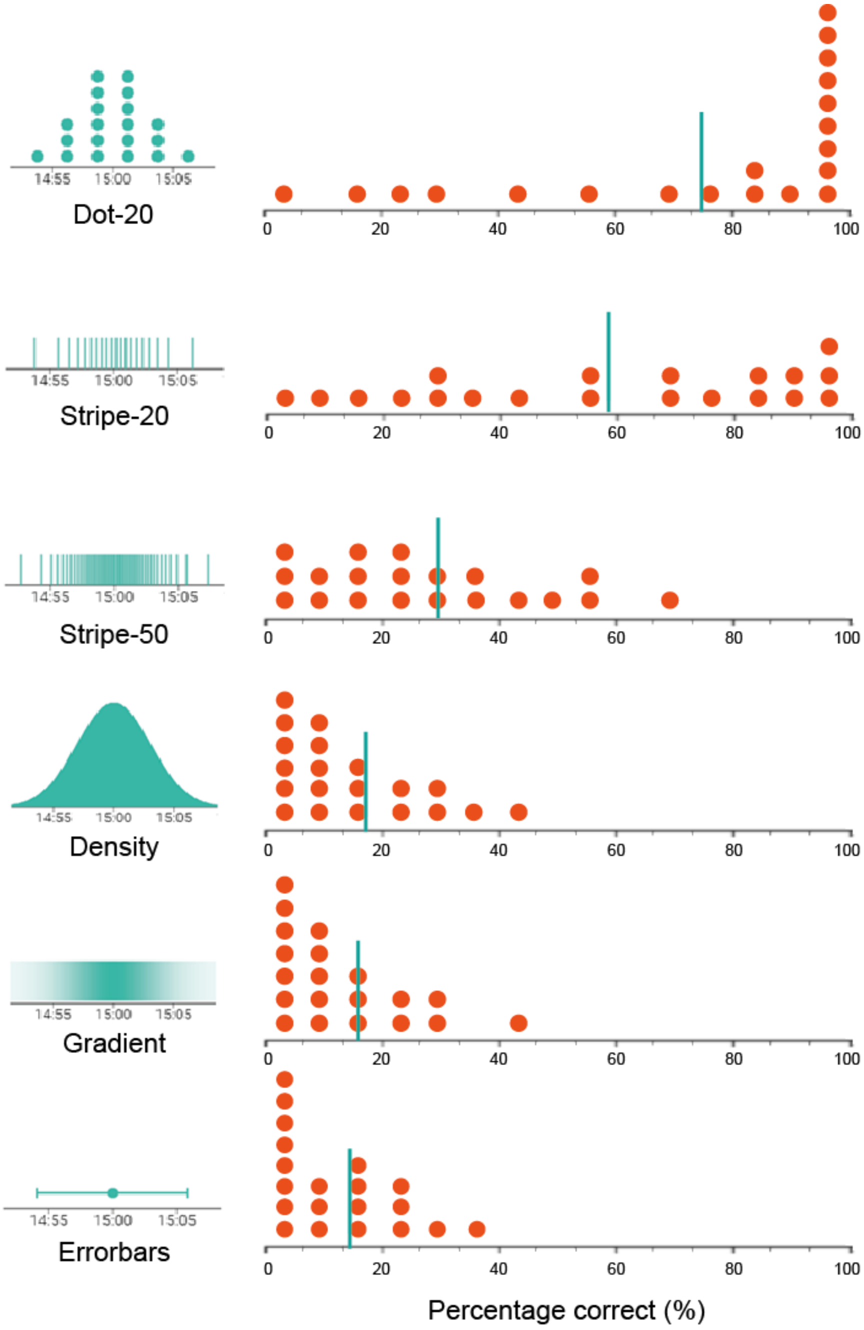


Figure 11. Average percentages correct on questions 2 and 3 per visualization type. Depicted by a Dot-20 visualization, meaning that one dot represents the performance level of 5% of the participants. The vertical stripes indicate the average percentage correct of all participants for that visualization type.

Comparing visualization types based on precision

To evaluate precision, a log transformation was performed on the estimated p's and the true p's. Then, error sizes were calculated by subtracting logit(*true p*) of logit(*estimated p*). The s-shaped function logit transforms probabilities into log-odds, which simplifies the analysis of probabilities (Kay et al., 2016). Figure 12 demonstrates the densities of those error sizes per visualization type, which gives insight in how precise people were in estimating probabilities. Error size densities from before the log transformation can be found in Appendix C. The dashed line indicates where error size is equal to zero. The order of visualizations in the graph (top down) is determined by their ranking on correctness. The narrower the frequency distribution, the lower the variance, the more precise participants were at estimating probabilities (Kay et al., 2016). The narrow, peaked distribution of Dot-20 indicates that participants' estimates were the most precise in that condition. The bottom four visualizations have wider and more diffuse distributions that indicate more variance and less precise estimations.

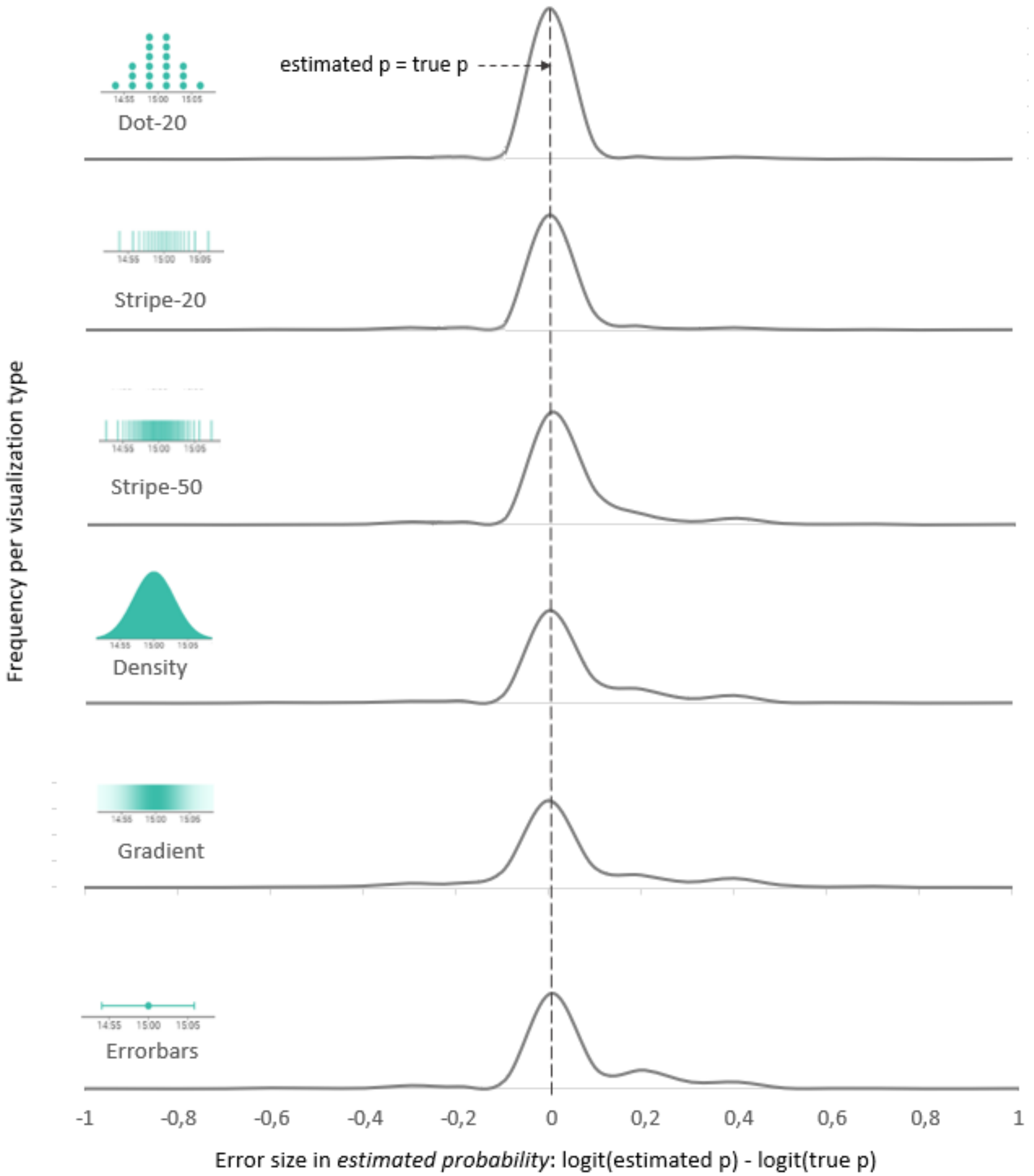


Figure 12. Precision (error size variability) per visualization type.

User characteristics

Correlating general task performance (Q2 & Q3) with user characteristics

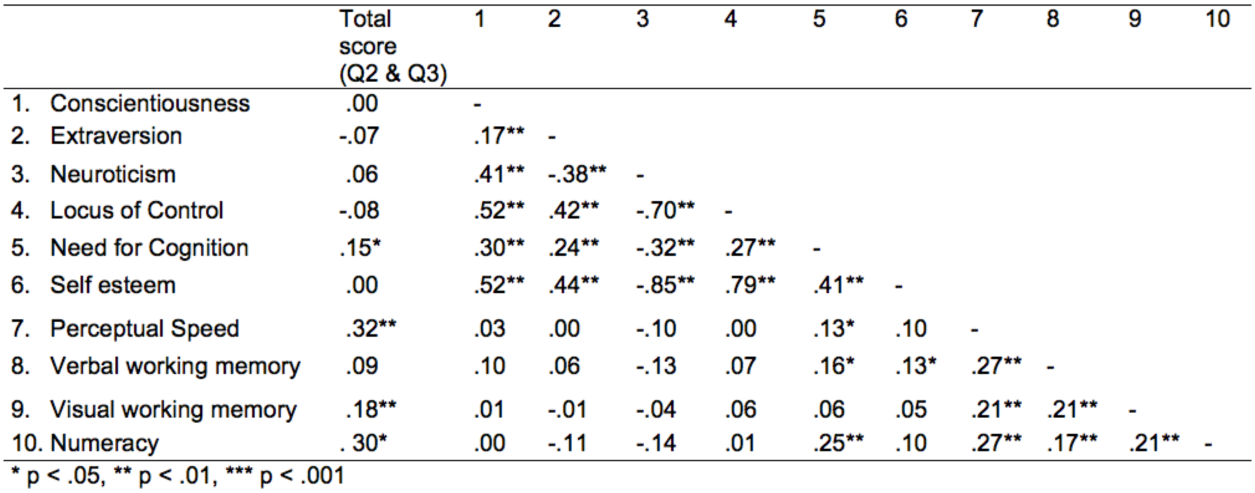
To assess the size and direction of the relationship between task performance on the graph task (measured in percentage correct on question 2 and 3) and scores on the set of psychometric measures that measure user characteristics, bivariate correlation coefficients were calculated. Statistical significance is reported at the 0.05 level. Prior to calculating these coefficients, normality, linearity and homoscedasticity were assessed, and found to be largely unsupported.

The assumption of normality is tested by Shapiro-Wilk statistic, which indicates that the data is not normally distributed when significance is below .05. Here, only *Conscientiousness* (*W* is .990, Sig = .104) and *Neuroticism* (*W* is .990, Sig = .104) do not violate the normality assumption. A visual inspection of the normal Q-Q and detrended Q-Q plots for each variable as an alternative to evaluate normality shows that *Extraversion* and *Expertise* do not violate the normality assumption either. All other variables suggest that the assumption of normality is violated. In that case, Spearman's Rho or Kendall's Tau-B are considered instead of Pearson's product-moment. Similarly, visually inspecting the scatterplots of all user characteristics against percentage correct on question 2 and question 3 in the graph task shows that the relationship between these variables was not linear or heteroscedastic. Therefore, Spearman's Rho is used.

The bivariate Spearman's rho correlations between the user characteristics scores and accuracy on Q2 and Q3 are reported in Table 6.

Table 6.

Spearman's Rho correlations between user characteristics and general task performance on question 2 and 3 (column 1). The numbers in the columns correspond to the numbered characteristics in the rows.



	Total score (Q2 & Q3)	1	2	3	4	5	6	7	8	9	10
1. Conscientiousness	.00	-									
2. Extraversion	-.07	.17**	-								
3. Neuroticism	.06	.41**	-.38**	-							
4. Locus of Control	-.08	.52**	.42**	-.70**	-						
5. Need for Cognition	.15*	.30**	.24**	-.32**	.27**	-					
6. Self esteem	.00	.52**	.44**	-.85**	.79**	.41**	-				
7. Perceptual Speed	.32**	.03	.00	-.10	.00	.13*	.10	-			
8. Verbal working memory	.09	.10	.06	-.13	.07	.16*	.13*	.27**	-		
9. Visual working memory	.18**	.01	-.01	-.04	.06	.06	.05	.21**	.21**	-	
10. Numeracy	. 30*	.00	-.11	-.14	.01	.25**	.10	.27**	.17**	.21**	-

* p < .05, ** p < .01, *** p < .001

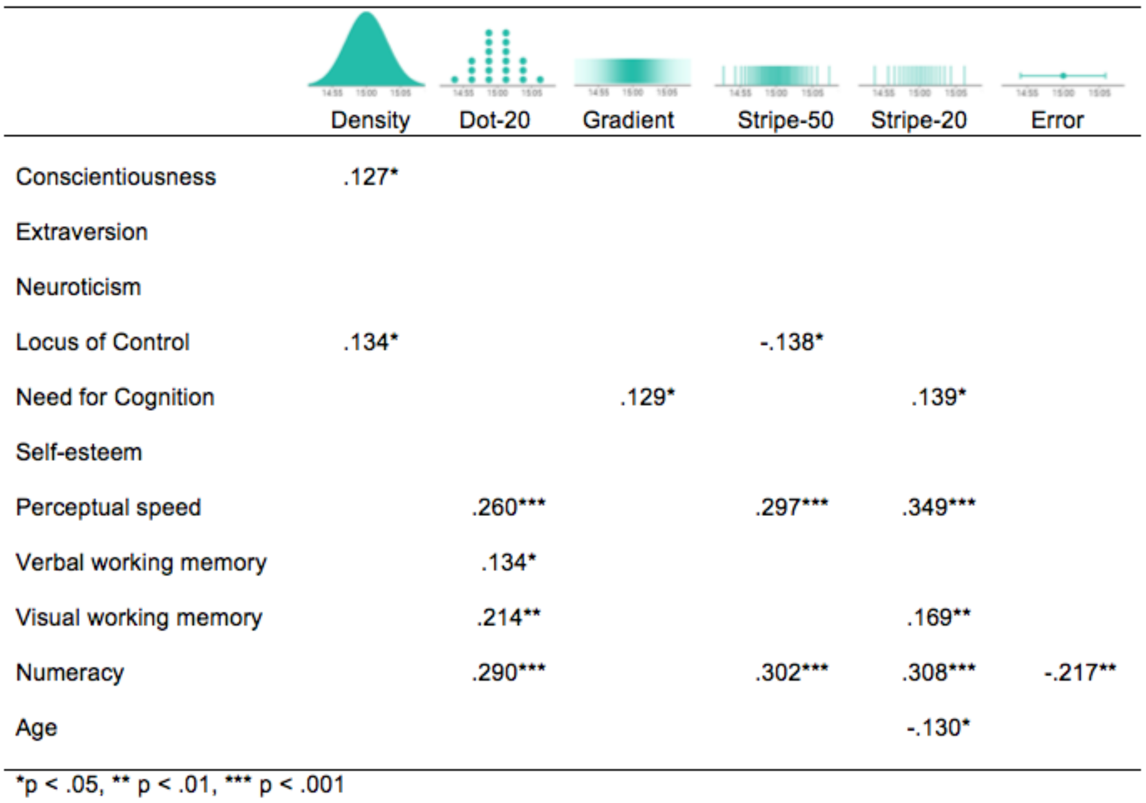
The first column shows that the personality trait Need for Cognition and the cognitive measures Perceptual Speed, Visual Working Memory and Numeracy prove to have a significant relation with task performance throughout all visualizations types. The rest of table shows that personality traits and cognitive measures show mutual significant correlations.

Correlating task performance per visualization type with user characteristics

Bivariate correlation coefficients were also calculated to assess the size and direction of the relationships between task performance on the graph task for each visualization type apart (measured in percentage correct for question 2 and 3) and scores on the set of psychometric measures that measure user characteristics. Because of the violated assumptions as described above, again, Spearman's Rho is used. Significant correlation coefficients are reported in Table 7.

Table 7.

Significant Spearman's Rho correlations between user characteristics and task performance on Q2 and Q3 per visualization type. Only significant correlation coefficients are reported.



	Density	Dot-20	Gradient	Stripe-50	Stripe-20	Error
Conscientiousness	.127*					
Extraversion						
Neuroticism						
Locus of Control	.134*			-.138*		
Need for Cognition			.129*		.139*	
Self-esteem						
Perceptual speed		.260***		.297***	.349***	
Verbal working memory		.134*				
Visual working memory		.214**			.169**	
Numeracy		.290***		.302***	.308***	-.217**
Age					-.130*	

*p < .05, ** p < .01, *** p < .001

To understand how these correlations manifest, the data was divided per user characteristic into 'high' and 'low' groups by performing a median split. The average percentages correct per split-group can be found in the median split table in Appendix D. After performing a median split, Dot-20 remains the best scoring visualization across all classifications.

Subjective measures

To evaluate how task performance (measured by correctness) is related to subjective measures, average rating were compared and correlation coefficients were calculated per visualization type. Figure 13 shows the average scores of the obtained ratings on a 5-point Likert scale in the post-questionnaire.

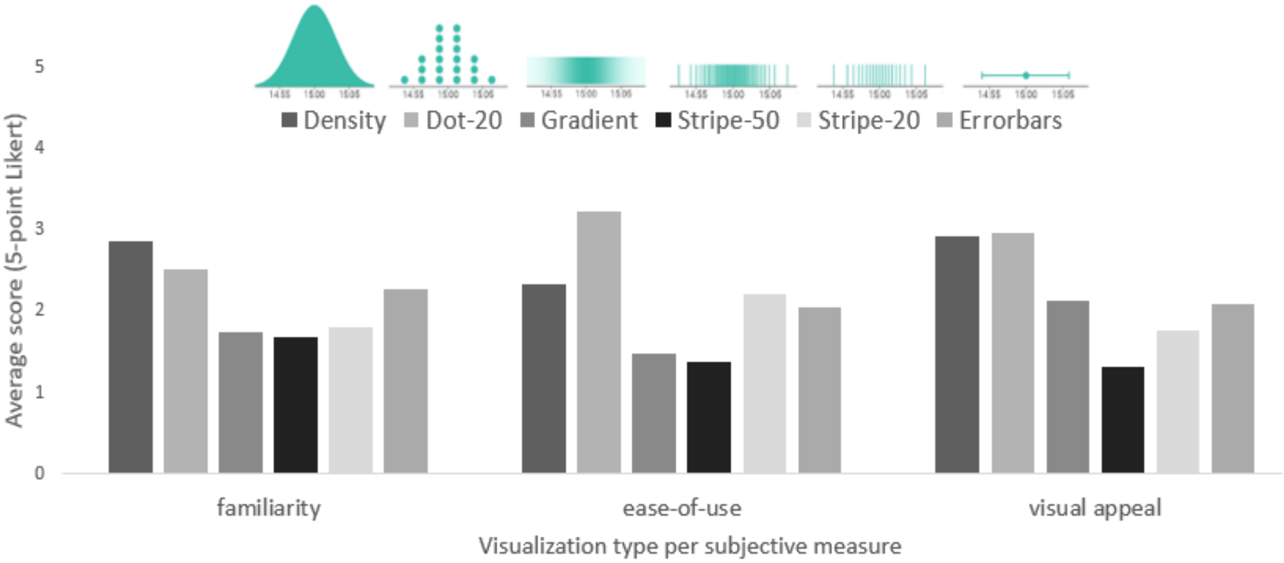
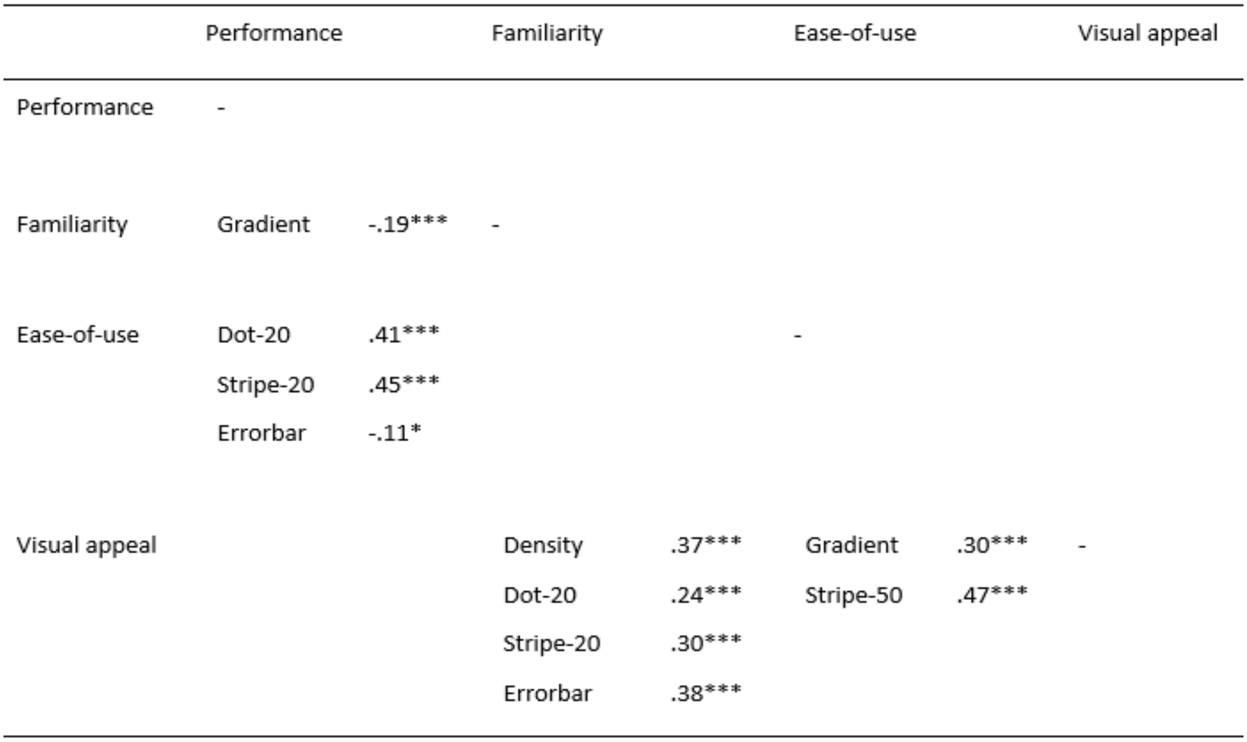


Figure 13. Average ratings on the subjective measures chart familiarity, ease-of-use, and visual appeal that were assessed in the post-questionnaire per visualization type.

The correlations that proved to be significant are reported in Table 8.

Table 8.

Correlation coefficients of the subjective measures (Kendall’s tau b)



	Performance		Familiarity		Ease-of-use		Visual appeal
Performance	-						
Familiarity	Gradient	-.19***	-				
Ease-of-use	Dot-20	.41***			-		
	Stripe-20	.45***					
	Errorbar	-.11*					
Visual appeal			Density	.37***	Gradient	.30***	-
			Dot-20	.24***	Stripe-50	.47***	
			Stripe-20	.30***			
			Errorbar	.38***			

The results will be further interpreted in the following Discussion.

4. Discussion

Summary

The goal of this study was to identify a visualization type best suited for conveying probability information to the general public, while also exploring if uncertainty visualization effectiveness might differ across user types, based on several cognitive- and personality characteristics. Six visualization types were selected and implemented in an online study with a car navigation system use-case. Participants were asked to judge best estimate of arrival time and to report two types of probability estimates (*later than* and *within a range*) based on a visualization that resembled a timeline. Several filter criteria caused 78 of 323 participants to be filtered out of the dataset, resulting in a sample size of 245 participants. The first research question focused on how accuracy and precision in probability estimates compared across six visual representations of statistical uncertainty. Results of the online graph task showed a significant main effect of visualization type, which indicates that the ability to estimate probabilities based on a visualization is influenced by the way the probability distribution is visually represented. The visualization type Dot-20 resulted in responses that were both high in accuracy and precision, followed up by Stripe-20 in both measures. Stripe-50 came in third and resulted in a significantly higher percentage of correct scores than the continuous types Density, Gradient and Errorbars, but showed precision patterns that were no better than those bottom three visualizations. In addition to comparing different types of visualizations, the influence of several user characteristics on task performance was investigated. Especially cognitive measures proved to be significantly related to task performance. The research questions and corresponding results will be addressed in more detail in the following.

Research questions

RQ1 How do accuracy and precision in probability estimates compare across six visual representations of statistical uncertainty?

The findings of the current study are in line with the expectancy that probability estimates are the most accurate and correct when based on discrete outcome plots with few enough outcomes to benefit from subitizing. The visualization type Dot-20 that was adopted from the study of Kay et al. (2016) proved its excellence in the current study. Followed up by the newly introduced Stripe-20 and the from Kay et al. (2016) adopted Stripe-50. Dot-20 and Stripe-20 resulted in probability estimates that were both high in accuracy and precision. Although significantly more people manage to be correct with Stripe-50 than with Density, Gradient and Errorbars, when looking at precision, Stripe-50 shows the same error size patterns as the three continuous encodings. It seems that some people take the effort of counting the stripes, which results in being correct, whereas other people handle Stripe-50 as if it were a continuous visualization, which results in a greater variety of error sizes. As expected, probability estimates that are based on the visualization Errorbars are significantly less often correct than with the other visualizations. In line with literature, Errorbars do perform poorly.

RQ2: Is visualization effectiveness influenced by user characteristics? Are some visualizations better suited for specific user types?

To analyze if variation in task performance can be attributed to variation in user characteristics scores, it was planned to perform a regression analysis. However, since the majority of the statistical assumptions were violated, a regression analysis would not be able to draw accurate conclusion about the current dataset. Instead, parametric correlation coefficients were calculated to see how user characteristic scores are related to task performance in general and for visualization types specific.

Cognitive measures

In line with literature, cognitive measures proved to be significantly related to task performance. As hypothesized, users with high perceptual speed perform better across all visualizations. Visual Working Memory and Numeracy also proved to be significantly correlated with *overall* task performance. These correlations between user characteristics scores and estimation accuracy across all visualizations indicate which characteristics are related to the general ability to estimate probability. Given the superior effectiveness of visualization type Dot-20, which will be furtherly discussed in the following, the correlations with task performance on Dot-20 will be highlighted. Perceptual Speed, Visual working memory, Verbal working memory and Numeracy are significantly positively correlated with the performance with Dot-20. This indicates that the lower the levels of these cognitive measures, the worse users are estimating probabilities based on Dot-20. Implications of these results will be discussed later this section.

Personality

The personality trait Need for Cognition is significantly correlated with overall task performance, which can be seen as a representation of the general ability of estimating probabilities. The personality trait Need for Cognition refers to an individual's tendency to engage in effortful cognitive activities. Since the experiment was quite demanding in size and length, the correlation might arise from a participant's inherent willingness or even enjoyment to persist in an effortful cognitive task.

In contrast to earlier research, no other significant correlations or differences based on personality traits were found. A possible explanation is the lack of time-pressure in the current experiment. It is suggested in literature that the impact of individual differences might be more pronounced in time-based tasks (Toker et al., 2012). In a preliminary study, participants were asked to respond as fast as possible during the online visualization task and results showed strong effects of the personality trait extraversion on task performance (Venrooij, 2018). For follow-up research, time pressure could be added to the experiment.

Personalizing visualization type

The visualization type Dot-20 resulted in probability estimates that were significantly more accurate than all other visualizations *and* the most precise, hence the most effective. The excellence of Dot-20 even proved to be robust across all user classifications. The divisions into ‘low’ and ‘high’ level groups per characteristic (Appendix D) showed that, even though differences in task performance between user-groups and visualization types exist, Dot-20 remains by far the best scoring visualization for every user type. Thus, adaptation or personalization in terms of visualization type based on user characteristics would be irrelevant.

Even when looking at an individual level, taken together the probability estimates for question 2 and 3, 78.4% of all participants achieved their best results in terms of percentages correct when using the Dot-20. This means that 21.6% of all participants benefited more from other visualizations. If everyone would be given the Dot-20 only, 73.7% of all questions would be estimated correctly on average. If personalization would be provided in terms of visualization type and on an individual level, in other words, if each person would be given the visualization type that served him/her best, the average percentage correct would increase with 5.1% compared to the average percentage correct that would be yielded if every person would work with the visualization Dot-20 only.

Given the superiority of the Dot-20 across all classifications based on user characteristics, together with the finding that personalization in terms of visualization type on individual level would only yield a 5.1% increment in percentages correct *and* the intended purpose of this study to identify a widely supported and effective visual representation of data uncertainty, this study suggests that Dot-20 would be the best suited visualization for every audience.

RQ3: How is task performance (measured in percentage correct) related to subjective measures like chart familiarity and ratings of visual appeal and ease-of-use?

For Dot-20, Stripe-20 and Errorbars, task performance (measured in percentage correct responses) was significantly correlated with the subjective measure *ease-of-use*. The correlation coefficients of the two former visualizations were positive, which confirms that these two top performing visualizations were indeed easy to use accurately. In contrast, the correlation coefficient for Errorbar on performance and ease-of-use was negative, but nonetheless in line with literature, where error bars are often associated with misinterpretation. This correlation suggests that people feel like it is ease to use, while they are actually interpreting it wrongly and performing poorly.

As hypothesized, chart familiarity appears to be unrelated to task performance for five out of six visualizations. Except for a negative correlation for the visualization type Gradient, no significant correlations were found between self-reported familiarity and task performance with that specific visualization type. The single correlation suggests that users who reported that they had seen or used the visualization type Gradient before, did not perform well when using it and vice versa. There is no obvious explanation for this effect, but it does indicate that this poorly performing visualization type is not well understood.

Significant correlations between chart familiarity and visual appeal were found for the visualization types Density, Dot-20, Stripe-20 and Errorbars. Thus, for four out of six visualization types, the hypothesis that users with higher familiarity with a certain visualization rate that specific visualization higher in visual appeal than users with low familiarity is confirmed. This effect could be explained by the mere exposure effect; a psychological phenomenon that suggests that repeated exposure increases familiarity and that people tend to develop a preference for things or people that are more familiar to them than others (Falkenbach, Schaab, Pfau, Ryfa & Birkan, 2013).

A positive correlation between visual appeal and ease-of-use was found for Stripe-50 and Gradient. As shown in Figure 13 (p.34), Stripe-50 was rated the lowest of all visualizations on both measures. Although Stripe-50 *performs* significantly better than Density, Gradient and Errorbar in terms of accuracy, it is not preferred in terms of ease-of-use and visual appeal. As shown in Table 4 (p.27), the average reaction times associated with Stripe-50 responses are slower than the average reaction times of other visualizations. The average reaction time of Stripe-50 is 2,1 seconds slower than the question average of Q2 and 2,5 seconds slower than the question average of Q3. A plausible explanation is that people take their time to count the stripes, regardless of the relative many discrete outcomes, to then calculate the associated probability. This results in a reasonable percentage correct, but might be perceived as an inefficient hassle. This suggestion endorses the recommendation of Kay et al. (2016) to use discrete outcome plots with *few enough* outcomes to benefit from subitizing. Stripe-50 may perform reasonably well thanks to its discrete nature, but has too many outcomes to truly benefit from subitizing. Again, adding time pressure to a follow-up experiment could put Stripe-50 to the test and possibly undermine its success.

Considerations

Consideration implementation: Discrete outcomes & data loss

As noted earlier, the findings of the current study are in line with the study of Kay et al. (2016), who recommended discrete outcome plots with few enough outcomes to benefit from subitizing, like their Dot-20 that was adopted in the current study. The top performance of Dot-20, Stripe-20 and Stripe-50 confirms that converting the continuous probability range into a discrete visualization appears to benefit the ability to estimate probability as it enables quick and accurate information processing. However, the cost of the summarizing nature of these visualizations is data loss. Given the robust and positive effect of the few-outcome, discrete visualization types; data reduction seems to be beneficial. Still, it should be kept in mind that in some contexts data reduction equals costly data loss. The context of the implementation, the goal of the visualization, and the goal of the user should be evaluated in order to decide if the Dot-20 or Stripe-20 are suitable for the specific situation.

Consideration further research: What explains performance in more challenging conditions?

By comparing six visualizations that represents statistical data uncertainty, the goal of this study was to identify the one(s) that are best suited for conveying probability information to the general public, while also exploring if uncertainty visualization effectiveness might differ across user types, based on several cognitive- and personality characteristics. With this intended goal, analysis focused on the best performing visualizations that yielded high accuracy and correctness. The visualizations Dot-20 and Stripe-20 have proved to be relatively easy to use and suited for a broad audience. Therefore it might be less meaningful to know what kind of people excel when using these visualizations, since almost everybody can achieve high performance. The wheat is separated from the chaff when using the apparent harder to use visualizations, like the Gradient and Errorbar. Working with these visualizations appears to be more challenging, since significantly less people are able to estimate probabilities correctly when based on those visualizations. Research from Carenini et al. (2014) suggests that task performance depends on cognitive abilities more heavily as task complexity increases. Since less people are able to extract accurate information from those representations, it could be interesting to formulate a research questions that focuses on what user characteristics explain for high performance in hard conditions.

Limitation: Unknown test context

Using an online research platform makes testing with a large sample in a short period of time possible. However, it should be noted that the context in which participants respond to the task request is for the most part unknown. Here, context includes the physical and mental state the participants are in, possible distractions they face and the quality of their technological equipment. As an example, varying display types and different lighting conditions may influence the functionality of the gradient visualization, possibly making it less desirable (Tak, Toet & van Erp, 2014). Likewise, it remains unknown why some subjects perform poorly in a study. The goal of the task might have been unclear and misinterpreted or underperformance could have been caused by a misunderstood question or a failure of the visualization (Hullman, 2016). Unknown test circumstances makes it difficult to know if everyone is making a serious attempt . In order to motivate participants to take the task seriously, a bonus was provided based on responses. To check for the level of attention payed to the personality questionnaires, trick questions were built in. Participants that responded incorrectly to those trick questions, or showed repetitive answer patterns during the graph task, extreme reaction times or unreliably low scores on either the graph task or cognitive tasks were removed from the dataset and did not receive the bonus. Providing an incentive has shown to contribute to the reliability of data acquired in online studies and is especially appropriate for perception and cognition studies with responses that can easily be checked against a ground truth (Kosara & Ziemkiewicz, 2010). Filtering the data as described above was the final component of cleaning the data. In an online study, it is hard to guarantee the quality of the data as it tricky to decide what is right and what is wrong. Presumably, making a serious attempt is related to task performance in both the cognitive tasks as the graph task. To avoid invalid correlations that solely originate from the level of motivation a participant shows, filters were applied to all components of the experiment. In future work, it could be considered to shorten the experiment. On average, participants spent 56.8 minutes completing the whole experiment. It is plausible that a shorter task is less demanding and will cause a smaller variation in persistence and thus performance, which may benefit the validity and reliability of the data.

Limitation: validity of 'effectiveness'

In the current study, visualization effectiveness is defined as successful in enabling the quick extraction of accurate information (Kennedy, Hill, Allen & Kirk, 2016). Although this definition does not include the term intuitive, intuitiveness is a concept that is often associated with information visualization effectiveness in literature (Friedman, 2008; Chen, 2017). According to the dictionary, intuition is the ability to acquire knowledge without proof, evidence, or conscious reasoning, or without understanding how the knowledge was acquired. In the current study, participants are instructed with the notion that the discrete visualizations (Dot-20, Stripe-20 and Stripe-50) have a certain amount of symbols and that hence every symbol represents a fixed amount of probability. Unsurprisingly, the discrete visualizations perform well. Although the possibility to simply count is their strength, it is not 'without conscious reasoning'. The deviating average reaction time of Stripe-50 from the other visualizations for Q2 and Q3 suggests that people do take the time to count. It raises the question if the visualization is only 'effective' in the current task according to our current definition and standards. Although Dot-20 yields accurate, correct and fast responses *and* gets rated the highest on average on both ease-of-use and visual appeal, the limitation above could be considered for future research. In future work, the explicit instruction could be left out or used as a manipulation to measure the effect of instructing. In addition, time pressure could be added to the experiment, to encourage people to interpret the visualization in a glance, as they would do when dealing with an actual car navigation system.

Implementations

The acquired knowledge about data visualizations can guide interface designers in how to divide their attention and budget to where the most profit in performance can be yielded. The found effect of visualization type on performance suggests that the design of a chart is a fruitful focus. When there is room for personalization, this study suggests that a low-level audience in terms of cognitive abilities would perform significantly worse when dealing with statistical uncertainty and can presumably benefit from some extra guidance to help them get to the desired or acceptable level of performance. If one knows that the target group of the interface consists of mostly people with high levels of numeracy and perceptual speed, performance would be more stable, even across visualization types, hence some costs can then be spared.

These findings can also contribute to the development of real-time user-adaptive systems. Evidence from a large body of research shows that user characteristics can significantly influence performance when interacting with information visualizations and that information about a user's cognitive abilities, personality and chart expertise can even be used to predict visualization effectiveness. At a growing rate, people with all kinds of abilities and backgrounds need to make decisions based on digital data. The growing demand for personalization in addition to the evidence that visualization effectiveness can depend on individual user differences triggers the need for user-adaptive visualizations (Toker et al., 2016). User adaptive visualizations are "visualizations that can customize the interaction to support users according to their individual needs." (Toker et al., 2016, p.17). The goal is to develop design requirements for expert systems and to eventually design user adaptive visualization systems that are able to adapt to their unique user in realtime (Green & Fisher, 2010; Toker, Conati, Carenini & Haraty, 2012). Studies show that it is possible to infer user differences from a user's eye gaze behaviour during the interaction with information visualizations (Steichen, Carenini & Conati, 2013; Toker, Conati, Steichen & Carenini, 2013; Toker & Conati, 2014; Conati & Gingerich, 2015). By doing so, user characteristics can be predicted from eye tracking data that is obtained while users perform simple visualization tasks. User characteristics data and associated eye tracking data is used to train and evaluate machine learning models that can reliably predict relevant user characteristics and classify user types in order to adapt in real-time (Toker, 2016).

Although this body of research is promising, the visualizations used in these studies do not include representations of data uncertainty (mostly vertical bar graphs without error bars). As it was argued in the beginning of this study, providing information about uncertainty enables users to make better, more nuanced decisions and it increases their trust in the data. The current study complements the existing body of knowledge as it was able to reproduce some of the findings on the relation of user characteristics with task performance, while offering a new visualization type that is able to effectively convey statistical data uncertainty. Earlier work focuses on two possible forms of adaptation: selecting different visualizations for different users, and providing only some users with additional automatic support, to benefit their performance when inspecting a given visualization (Toker et al., 2012). Based on the findings of the current study, personalizing visualization type would not be beneficial since Dot-20 performs best across all user types. However, while given the Dot-20 type, the user may receive guidance or clarifications from the system if the system identifies this user as low on a trait or ability that has proven to be a predictor for success in using the current visualization to ensure effective visualization processing.

Conclusion

As the size and the complexity of datasets *and* the use of data visualizations continues to grow, it is crucial to strive for standard inclusion of uncertainty measures in visual representations. Earlier work has shown that providing uncertainty information can improve decision-making, although the advantages of including uncertainty critically depend on how it is communicated. The findings of this study suggest that probability distributions can best be conveyed by discrete visualizations with few enough outcomes to benefit from subitizing, as it proves to result in significantly more accurate and more precise probability estimations. Task performance differs across users with different levels of cognitive abilities and personality traits, but it can be concluded that visualization type has a much greater impact on performance than individual differences have. This suggests that, when designing an interface with an aim for high performance, it is more effective to focus on the graphic design of a chart than to put effort in personalization. In situations where personalization is wanted, the results of the current study suggest that providing extra guidance to users with low levels of perceptual speed, numeracy, visual- and verbal working memory capacity would be the most beneficial, given their considerable potential to gain in performance level.

References

Aerts, J. C., Clarke, K. C., & Keuper, A. D. (2003). Testing popular visualization techniques for representing model uncertainty. *Cartography and Geographic Information Science*, *30*(3), 249-261.

Allen, B. (2000). Individual differences and the conundrums of user-centered design: Two experiments. Journal of the Association for Information Science and Technology, 51(6), 508-520.

Askew, M., Rhodes, V., Brown, M., William, D., Johnson, D. (1997). *Effective teachers of numeracy*. London: King's College London.

Baddeley, A. D. (1986). Working memory. New York: Clarendon Press/Oxford University Press.

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological methods*, *10*(4), 389.

Blascovich, J., & Tomaka, J. (1991). Measures of self-esteem. *Measures of personality and social psychological attitudes*, *1*, 115-160.

Blenkinsop, S., Fisher, P., Bastin, L., & Wood, J. (2000). Evaluating the perception of uncertainty in alternative visualization strategies. *Cartographica: The International Journal for Geographic Information and Geovisualization*, *37*(1), 1-14.

Bonneau, G. P., Hege, H. C., Johnson, C. R., Oliveira, M. M., Potter, K., Rheingans, P., & Schultz, T. (2014). Overview and state-of-the-art of uncertainty visualization. In *Scientific Visualization* (pp. 3-27). Springer, London.

Carenini, G., Conati, C., Hoque, E., Steichen, B., Toker, D., & Enns, J. (2014, April). Highlighting interventions and user differences: informing adaptive information visualization support. In Proceedings of the 32nd annual ACM conference on Human factors in computing systems (pp. 1835-1844). ACM.

Chen, H. M. (2017). An Overview of Information Visualization. *Library Technology Reports*, *53*(3), 5.

Conati, C., & Maclaren, H. (2008, May). Exploring the role of individual differences in information visualization. In Proceedings of the working conference on Advanced visual interfaces (pp. 199-206). ACM.

Daken, A., Dogruel, E., Grimes, J., Lam, M., and Lotze, T. (2008). Uncertainty for the Novice; Fuzz: A Visualization Toolkit.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. The Mini-IPIP scales: Tiny-yet-effective measures of the Big Five factors of personality. Psychological Assessment, (2006), 192-203.

Dragicevic, P. (2016). Fair statistical communication in HCI. In *Modern Statistical Methods for HCI* (pp.

291-330). Springer, Cham.

Ekstrom, R.B., French, J.W., Harman, H.H. Kit of factor-referenced cognitive tests. Educational Testing Service, Princeton, NJ, (1976).

Evans, B. J. (1997). Dynamic display of spatial data-reliability: Does it benefit the map user?. *Computers & Geosciences*, *23*(4), 409-422.

Falkenbach, K., Schaab, G., Pfau, O., Ryfa, M., & Birkan, B. (2013). Mere Exposure Effect.

Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten*. Analytics Press.

Few, S., & Edge, P. (2007). Data visualization: past, present, and future. *IBM Cognos Innovation Center*.

Friedman, V. (2008). Data Visualization and Infographics| Smashing Magazine.

Fukuda, K., & Vogel, E.K. Human variation in overriding attentional capture. Journal of Neuroscience, (2009), 8726-8733.

Gershon, N. (1998). Visualization of an imperfect world. *IEEE Computer Graphics and Applications*, *18*(4), 43-45.

Green, T. M., & Fisher, B. (2010, October). Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on (pp. 203-210). IEEE.

Greis, M., Joshi, A., Singer, K., Schmidt, A., & Machulla, T. (2018, April). Uncertainty Visualization Influences how Humans Aggregate Discrepant Information. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 505). ACM.

Greis, M., Ohler, T., Henze, N., & Schmidt, A. (2015, September). Investigating representation alternatives for communicating uncertainty to non-experts. In *Human-Computer Interaction* (pp. 256-263). Springer, Cham.

Greis, M., Schuff, H., Kleiner, M., Henze, N., & Schmidt, A. (2017). Input Controls for Entering Uncertain Data: Probability Distribution Sliders. *Proceedings of the ACM on Human-Computer Interaction*, *1*(1), 3.

Griethe, H., & Schumann, H. (2006, March). The visualization of uncertain data: Methods and problems. In *SimVis* (pp. 143-156).

Gschwandtner, T., Bögl, M., Federico, P., & Miksch, S. (2016). Visual encodings of temporal uncertainty: A comparative user study. *IEEE transactions on visualization and computer graphics*, *22*(1), 539-548.

Hullman, J. (2016, October). Why evaluating uncertainty visualization is error prone. In *Proceedings of* *the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization* (pp. 143-151). ACM.

Ibrekk, H., & Morgan, M. G. (1987). Graphical communication of uncertain quantities to nontechnical people. *Risk analysis*, *7*(4), 519-529.

International Organization for Standardization. (1994). *ISO 5725-2: 1994: Accuracy (Trueness and Precision) of Measurement Methods and Results-Part 2: Methods for the Determination of Repeatability and Reproducibility*. International Organization for Standardization.

Jackson, C. H. (2008). Displaying uncertainty with shading. *The American Statistician*, *62*(4), 340-347.

Joslyn, S. L., & LeClerc, J. E. (2012). Uncertainty forecasts improve weather-related decisions and attenuate the effects of forecast error. *Journal of experimental psychology: applied*, *18*(1), 126.

Joslyn, S., & LeClerc, J. (2013). Decisions with uncertainty: the glass half full. *Current Directions in Psychological Science*, *22*(4), 308-315.

Kay, M. J. S. (2016). *Designing for User-facing Uncertainty in Everyday Sensing and Prediction* (Doctoral dissertation).

Kay, M., Kola, T., Hullman, J. R., & Munson, S. A. (2016, May). When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5092-5103). ACM.

Kennedy, H., Hill, R. L., Allen, W., & Kirk, A. (2016). Engaging with (big) data visualizations: Factors that affect engagement and resulting new definitions of effectiveness. *First Monday*, *21*(11).

Kosara, R., & Ziemkiewicz, C. (2010, April). Do Mechanical Turks dream of square pie charts?. In *Proceedings of the 3rd BELIV'10 Workshop: Beyond time and errors: Novel evaluation methods for information visualization* (pp. 63-70). ACM.

Lapinski,A. L. S. (2009). *A strategy for uncertainty visualization design* (No. DRDC-ATLANTIC-TM-2009 151). Defence Research and Development Atlantic Dartmouth (Canada).

Levenson, H. (1981). Differentiating among internality, powerful others, and chance. In H. M. Lefcourt (Ed.), Research with the locus of control construct (Vol. 1, pp. 15–63). New York, NY: Academic Press.

Lohse, G. L. (1997). Models of graphical perception. Handbook of Human-Computer Interaction, 2, 107 135

MacEachren, A. M., Roth, R. E., O'Brien, J., Li, B., Swingley, D., & Gahegan, M. (2012). Visual semiotics & uncertainty visualization: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2496-2505.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, *5*(2), 110-141.

Morss, R. E., Demuth, J. L., & Lazo, J. K. (2008). Communicating uncertainty in weather forecasts: A survey of the US public. *Weather and forecasting*, *23*(5), 974-991.

Munzner T. Visualization Analysis & Design. CRC Press, (2015).

Olston, C., & Mackinlay, J. D. (2002). Visualizing data with bounded uncertainty. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on* (pp. 37-40). IEEE.

Pang, A. T., Wittenbrink, C. M., & Lodha, S. K. (1997). Approaches to uncertainty visualization. *The Visual Computer*, *13*(8), 370-390.

Patterson, R. E., Blaha, L. M., Grinstein, G. G., Liggett, K. K., Kaveney, D. E., Sheldon, K. C., & Moore, J. A. (2014). A human cognition framework for information visualization. *Computers & Graphics*, *42*, 42-58.

Roberts, B. W., Jackson, J. J., Fayard, J. V., Edmonds, G., & Meints, J. (2009). Conscientiousness.

Rosenberg, M. (1965). Rosenberg self-esteem scale (RSE). *Acceptance and commitment therapy. Measures package*, *61*, 52.

Roulston, M. S., Bolton, G. E., Kleit, A. N., & Sears-Collins, A. L. (2006). A laboratory study of the benefits of including uncertainty information in weather forecasts. *Weather and Forecasting*, *21*(1), 116-122.

Schapira, M. M., Nattinger, A. B., & McHorney, C. A. (2001). Frequency or probability? A qualitative study of risk communication formats used in health care. *Medical Decision Making*, *21*(6), 459-467.

Steichen, B., Carenini, G., & Conati, C. (2013, March). User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In Proceedings of the 2013 international conference on Intelligent user interfaces (pp.317-328). ACM.

Tak, S., Toet, A., & van Erp, J. (2014). The Perception of Visual Uncertainty Representation by Non Experts. *IEEE transactions on visualization and computer graphics*, *20*(6), 935-943.

Tak, S., Toet, A., & Van Erp, J. (2015). Public understanding of visual representations of uncertainty in temperature forecasts. *Journal of cognitive engineering and decision making*, *9*(3), 241-262.

Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., & Pavel, M. (2005, March). A typology for visualizing uncertainty. In *Visualization and Data Analysis 2005* (Vol. 5669, pp. 146-158). International Society for Optics and Photonics.

Toker, D., & Conati, C. (2014, July). Eye tracking to understand user differences in visualization processing with highlighting interventions. In International Conference on User Modeling, Adaptation, and Personalization (pp. 219-230). Springer, Cham.

Toker, D., Conati, C., Carenini, G., & Haraty, M. (2012). Towards adaptive information visualization: on the influence of user characteristics. User Modeling, Adaptation, and Personalization, 274-285.

Toker, D., Conati, C., Carenini, G., Munzner, T., & Enns, J. (2016). Exploring user-adaptive visualization techniques with Multi-Modal Documents.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2).

Turner, M. L., & Engle, Randall W. Is working memory capacity task dependent? Journal of Memory and Language, (1989), 127-154.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, *185*(4157), 1124-1131.

Venrooij, W., Toet, A., & van Erp, J. (2018, in press). Personal differences in chart effectiveness.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*(2), 198-212.

Wilkinson, L. (1999). Dot plots. *The American Statistician*, *53*(3), 276-281.

Wu, C. H., Parker, S. K., & De Jong, J. P. (2014). Need for cognition as an antecedent of individual innovation behavior. *Journal of Management*, *40*(6), 1511-1534.

Ziemkiewicz, C., Crouser, R. J., Yauilla, A. R., Su, S. L., Ribarsky, W., & Chang, R. (2011,October). How locus of control influences compatibility with visualization style. In Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on (pp. 81-90). IEEE.

Appendix A: Terminology explained

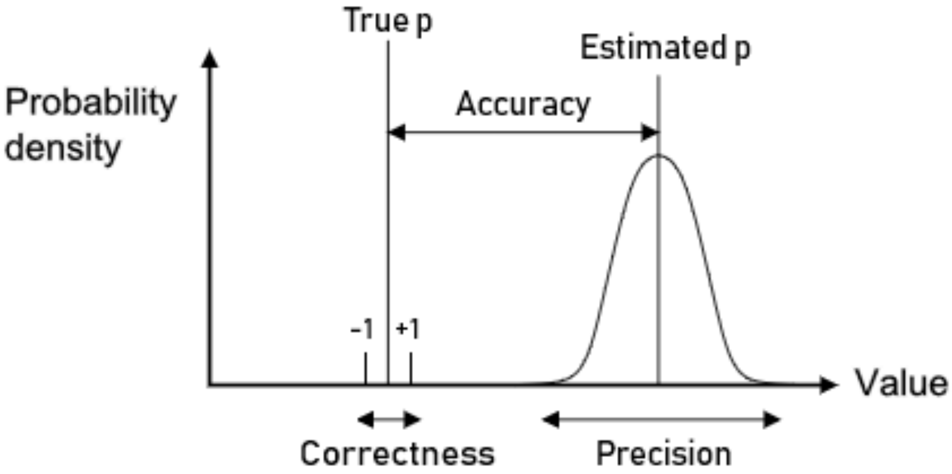


Figure 14. Terminology explained in a figure. Figure is adopted from ISO 5725-1 (1994) and adjusted to the terminology of the current study.

Table 9.

The term and concept 'accuracy' explained.

Concept	Accuracy
Definition	The proximity of a measurement to the true value (ISO 5725-1, 1994)
Operationalization	The difference between the probability estimate given by the participant (*estimated p*) and the true value (*true p*). This ground truth is based on the underlying probability distribution of the visualization.
Metric	Error size, the resulting value of *estimated p - true p*
Scale	Error size can range from -100 to 100. Probability estimates range from 0% to 100%, so the maximum absolute error size is 100. When error size is negative, one has underestimated the true value and when error size is positive, one has overestimated the true value.
Interpretation	The closer error size is to 0, the higher the accuracy, the better the performance.

Table 10.

The term and concept 'correctness' explained.

Concept	Correctness
Definition	The quality or state of being free from error (Oxford Dictionary, 2018). Being accurate.
Operationalization	A response is considered correct when error size ≤ 1 and incorrect when error size > 1. Correct if *estimated p* is equal to *true p* or 1 point off *true p*. (*estimated p* = *true p*; *estimated p* = (*true p* - 1); *estimated p* = (*true p* + 1)).
Metric	Error size
Scale	Binary scale: correct or incorrect.
Interpretation	Correctness is a binary form of accuracy used to compare average accuracy. For interpretation often transformed into *percentage* (%) correct (i.e. per participant, task type, visualization type etc.). The higher the correctness / percentage correct, the better the performance.

Table 11.

The term and concept 'precision' explained.

Concept	Precision
Definition	The closeness of agreement among a set of results (ISO, 2012).
Operationalization	The variability of the estimated p's
Metric	Error size distribution: the distribution of all resulting values of *estimated p* - *true p*.
Scale	Error size can range from -100 to 100, hence this is the range of the distribution.To plot and evaluate precision, a log transformation was performed on the estimated p's and the true p's. Then, error sizes were

	calculated by subtracting logit(*true p*) of logit(*estimated p*). The s-shaped function logit transforms probabilities into log-odds, which simplifies the analysis of probabilities (Kay et al., 2016). The resulting error size distribution ranges from log odds -1,5 to 1,5.
Interpretation	The narrower the error size distribution, the lower the variance, the more precise the estimates, the better the performance.

Table 12.

The term and concept ‘effectiveness‘ explained.

Concept	Effectiveness
Definition	Successful in enabling the quick extraction of accurate information (Kennedy et al., 2016). Effectiveness is used as an overarching term to characterize the general quality of a visualization type.
Operationalization	A visualization type is considered *effective* if the responses (the *estimated p’s*) based on that visualization type are high in accuracy, precision and correctness.
Metric	Accuracy, precision and correctness
Scale	See Table 9, 10 and 11.
Interpretation	There is no absolute interpretation for effectiveness. Rather it is used to judge relative effectiveness among the various visualization types.

Appendix B: Description of non-selected encoding

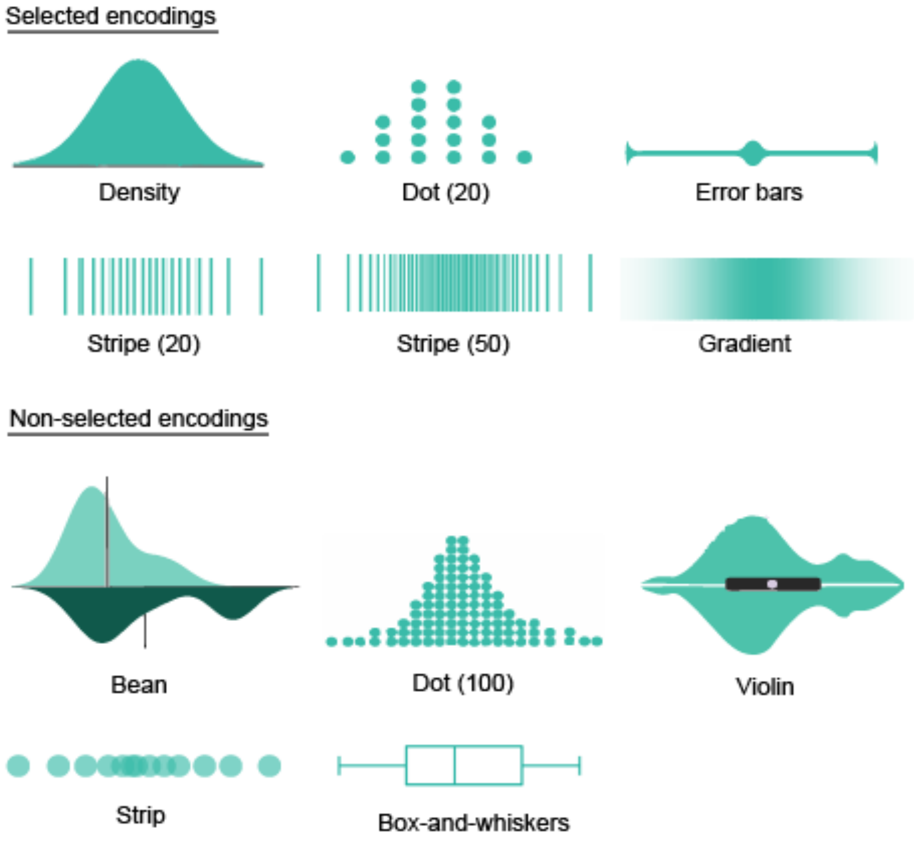


Figure 15. Overview of the visual encodings that were considered for the study (based on literature).

<u>Selected encodings</u>

Error bars

Error bars are frequently used, cap-tipped lines that serve as a graphical enhancement to display the uncertainty of the plotted data. They can be applied to scatter plots, dot plots, line graphs, and bar graphs. Despite their popularity, error bars received a lot of critique due to their severe shortcomings (Correll & Gleicher, 2014). Correll & Gleicher (2014) investigated the drawbacks of the standard encoding of mean and error and evaluated alternatives. They concluded that bar graphs with error bars suffer from two major biases: the *within-the-bar bias*, meaning that values within the bar are judged as likelier than values outside the bar, because the glyph of a bar provides a false metaphor of containment. Further, they suffer from *binary interpretation*; values are within the margins of error, or they are not (Correll & Gleicher, 2014). Cairo (2016) also argues that this all-or-nothing quality of error bars is the most obvious shortcoming of the encoding. Since error bars often represent a confidence interval, they work as a probability distribution following the theoretical idea that values closer to the upper and lower boundaries are less likely than values to the reported point estimate. These finer details about the probability distribution are hidden in error bars (Cairo, 2016). An empirical study with 473 respondents suggests that many prominent researchers have a poor understanding of how error bars relate to statistical significance (Belia, Fidler, Williams & Cumming, 2005). Moreover, the chosen significance cutoff value that indicates the size of the confidence interval is often arbitrary (Krusz, 2013). That is, if error bars are used to represent a confidence interval, which is not always the case. Dragicevic (2016) argues that researchers need to become more consistent and more clearly in indicating what error bars refer to. Error bars are ambiguous, since they are used to encode confidence intervals (ranging from 80% to 95%), standard deviation, and standard error (Kay, 2016). This lack of standardization makes it even for trained scientists hard to interpret the data, which can lead to incorrect conclusions (Dragicevic, 2016). To mitigate some of the problems associated with error bars, Correll & Gleicher (2014) proposed alternative encodings that are *visually symmetric* and *visually continuous*. Violin and gradient plots are example solutions. Error bars are included in the current study to check for the drawbacks described in earlier work.

Density plot

A density plot is a function graph of a probability distribution function and encodes the density as distance from the x-axis (Kay, 2016). The resulting curve (or *area*) provides a simple summary of the distributions shape and enables quick visual interference about the distribution of the data. To convey the probability density, the density plot relies on the width of an area. According to Mackinlay's (1986) ranking of effective visual encodings, the size of an area is considered to be quicker and more easily perceived than other retinal variables such as opacity, color, or texture. It enables the viewer to detect clusters or bumps within a distribution at first glance. The plot does not show precise numbers, nor does it show a measure of center (Few, 2015). The probability distribution and the mode (visually encoded by the maximum of the density) are intrinsic to each other in a density plot, making an explicit measure of center otiose.

Results of Greis, Ohler, Henze & Schmidt (2015) show that a density plot is the best way to communicate uncertain information to non-experts. In the study of Kay et al. (2016), the density plot results in accurate probability judgements and is also highly rated in terms of visual appeal. The violin plot, which is suggested as a superior alternative to error bars by Correll & Gleicher (2014), is based on the density plot and relies on the same principles to convey the shape of the distribution. Although the density plot lacks the recommended visual symmetry of the violin plot, it still mitigates error-bar issues by being visually continuous. In addition, the unilateral appearance of the density plot reduces issues with visual clutter in comparison to bilateral encodings such as the violin and bean plot. The density plot is included in the current study, because of its known effectiveness for communicating uncertainty and its visual simplicity compared to the violin and bean plot.

Dot plot

A dot plot shows discrete quantiles on a continuous scale using a dot or other symbol (Wilkinson, 1999). By stacking the dots, the plot shows the distribution of the data, but does not include any graphical description of summaries (Benjamini, 1988). However, the possibility to manipulate the amount of dots used to represent the data, can make summary statistics unnecessary. In a low-density dot plot of 20 dots (referred to as Dot-20), every dots represents 5% of the observations, while a dot in a high-density dot plot of 100 dots (Dot-100) represents 1% of the observations. Kay (2016) shows that both encodings have their own advantages. He finds that a Dot-20 allows the viewer to count the dots in the tails and body of the distribution, enabling quick, accurate judgements without any summary statistics. In Dot-100, counting is irrelevant; however, density is very well-resolved.

In the study of Kay et al. (2016), Dot-20 resulted in the most precise probability estimates and performed best across all conditions. Dot-100 performed very similarly to the density plot. Because of its superior performance in the study of Kay et al. (2016), the dot plot with 20 dots is included in the current study. In the interest of selecting six encodings that represent a wide range of possible trade-offs in visualization properties, the Dot-100 is dropped.

Stripe plot

The stripe plot is a variation on the strip plot; a one-dimensional scatter plot representing individual observations or probabilities using a dot, or in case of the stripe plot a stripe. Probability density is thus encoded by the density of vertical stripes in a region (Kay, 2016). They are especially useful for small batches of data and for comparing multiple distributions at once (Few, 2012). However, for large sets, a strip plot can easily suffer from overplotting: multiple points in the same location. This can either be solved by stacking the symbols that overlap, resulting in a dot plot or by adjusting the opacity of the symbol to its corresponding probability density, resulting in a gradient plot. When there are (too) many outcomes, a discrete plot like the stripe plot converges into a continuous encoding. Like dotplots are a discrete analog to a density plot, stripe plots can be considered the discrete analog to a gradient plot

(Kay, 2016).

In the study of Kay (2016), stripe plot-50 performed poorly, as it elicited in respondents' probability estimates that were the least precise of all conditions. Kay (2016) explains its underperformance by addressing the phenomenon of the stripe plot being converged and read like a continuous plot. However, this explanation does not resolve the issue, since Dot-100, which can be seen as the discrete analog to the density plot, did not suffer from this potential convergence as it performed well and similarly to the density plot in the study of Kay et al. (2016). In order to test the influence of the amount of outcomes displayed in a discrete encoding, a stripe plot with 20 stripes (Stripe-20) is introduced and the Stripe-50 of Kay (2016) will be adopted in the current study. Moreover, gradient plots, the actual continuous analog to the stripe plot (not included in the study of Kay et al. 2016)), are known to perform well in displaying statistical uncertainty (Correll & Gleicher, 2014; Gschwandtner et al, 2016). To investigate the effect of discrete and continuous encodings of continuous distributions on probability estimates, both the stripe plot and the gradient plot will be included in the current study.

Gradient plot

A gradient plot is a shaded horizontal bar glyph of fixed height and width, in which the probability density of the quantity at a point is encoded by opacity. The darker the shade, the higher the probability of the given estimate (Jackson, 2008). In contrast to error bars, which may give the false binary interpretation that all points within the whiskers are equally likely and that points outside the lines are impossible, a gradient plot gives a fuller description of the uncertainty surrounding the parameter estimate by representing the entire distribution in one dimension without terminating at a clear limit (Jackson, 2008). Although opacity is known to be a less effective way of visually encoding density than width and area (as used in violin and density plots) (Mackinlay, 1986), the purpose of the gradient plot is to indicate the shape of the distribution, rather than to allow precise determination of the value of each point (Jackson, 2008). Moreover, the fixed size of the gradient plot benefits perception as it reduces visual clutter. The visual variable 'fuzziness' is intuitively linked to uncertainty and is therefore a suitable visual metaphor for communicating uncertainty (Gherson, 1998; MacEachren et al., 2012)

Both the stripe and gradient plot are especially useful for accurately extracting the point of maximum probability density (Ibrekk & Morgan, 1987). Gradients are especially useful for representing statistical uncertainty (Gschwandtner et al., 2016) and can be used to communicate a wide variety of data sources with an underlying normal distribution (Tak, Toet & van Erp, 2014). Correll & Gleicher recommend the gradient plot as an alternative to error bars and box plots, as it is *visually symmetric* and *visually continuous*. Other work shows that the gradient plot is superior to violin plots and accumulated probability plots in communicating probabilities (Gschwandtner et al., 2016). The effectiveness of the gradient plot will be evaluated in the current study.

Non-selected encodings

Box-and-whiskers plot

John Tukey's box-and-whiskers plot (often referred to as the box plot) gives a five number summary of a batch of data, which consists of the largest, smallest, median, and upper and lower quartiles (Tukey, 1977). This graphical statistic summary makes the location, spread, skewness, and longtailed-ness of the data available with a quick glance (Benjamini, 1988). The box plot can tell a lot about a distribution, while still remain its simplicity (Few, 2015). In some cases, outliers are displayed as individual dots independently from the whiskers. However, this demands the binary decision whether an observation should be considered as an "outlier" and this can be quite arbitrary, especially in case of non-normal underlying distributions (Dragicevic, 2016; Kampstra, 2008). There have been suggested variations of the boxplot, in which density information is included and conveyed by the sides of the box (Benjamini, 1988), from which the violin plot is the most successful example (Hintze & Nelson, 1998). Last but not least, when using the boxplot to present information to others, it should be kept in mind that most people in the world have never learned how to read a boxplot and that it therefore might need a short explanation (Few, 2015). The box plot is not included in the current study, because it is visually clunkier than error bars and still suffers from some of the same problems (Krusz, 2013). Although the box plot conveys more information about the shape of the distribution than error bars, the need for prior knowledge does not make the box plot an intuitive and appropriate alternative to error bars.

Violin plot

A violin plot combines the common components of a box plot (i.e. upper and lower values, quartiles and the median) with density traces that are plotted symmetrically to both sides of the vertical box plot (Hintze & Nelson, 1998). The resulting single, symmetrical plot contains the basic summary statistics, while also providing an indication of the shape of the distribution including any existing clusters in data . This together makes the violin plots a valuable tool for data analysis and exploration (Hintze & Nelson, 1998). Like the density plot, the violin plot relies on the retinal variables area and width. These effective visual encodings enable the viewer to detect clusters or bumps within a distribution at first glance.

Correll & Gleicher (2014) recommend the violin plot to mitigate some of the problems associated with error bars, as it is *visually symmetric* and *visually continuous*. The researchers show that, despite its unfamiliarity, the violin plot offers performance advantages to a general audience. The violin plot is not included in the current study, since the selected density plot is able to convey sufficient information about the shape of the distribution without redundant visual clutter.

Bean plot

The bean plot looks like the violin plot, but replaces the interior box plot with lines representing individual observations. Since it relies on the same retinal variables as the violin plot, it has the same advantages with regards to visual inference. A *bean* plot combines a density trace of a distribution (the *pod*) with a one-dimensional scatter plot representing individual data points as small lines (the *beans*). To enable quick comparison, the overall average is drawn as a vertical line. The bean plot can be symmetric like a violin plot, but can also be used to display two subgroups simultaneously in a special asymmetric bean plot (Kampstra, 2008). This makes the bean plot a helpful tool for visually comparing multiple batches of data. Another added value of the bean plot in comparison to the violin plot is that it all individual data points are shown, which provides information about the number of observations in a group and makes outliers detectable. Hence, there is no need for dichotomous assumptions about outliers, as discussed with regards to the boxplot. The bean plot is not included in the current study, since the selected density plot is able to convey sufficient information about the shape of the distribution without redundant visual clutter.

Strip plot

A strip plot is a one-dimensional scatter plot representing individual observations using a dot or other symbol. They are especially useful for small batches of data and for comparing multiple distributions at once (Few, 2012). However, for large sets, a strip plot can easily suffer from overplotting: multiple points in the same location. There are several ways to solve this. With relatively few values it is an option to *jitter* the data points. Jittering is the act of repositioning points that overlap either horizontally or vertically, so that they're no longer on top of each other (Few, 2012). For larger datasets, points can be made transparent, which makes the overplotted areas denser in color, allowing to see variation in the number of values (Few, 2012). A third way to solve overplotting is to use symbol size to encode the frequency of an observation or data point (Kirk, 2016). Strip plots are not included in the current study, since they do not show the shape of a statistical distribution very well (Few, 2012). They are valuable in the comparison of small batches of observations.

Appendix C: Error size curve before log-transformation

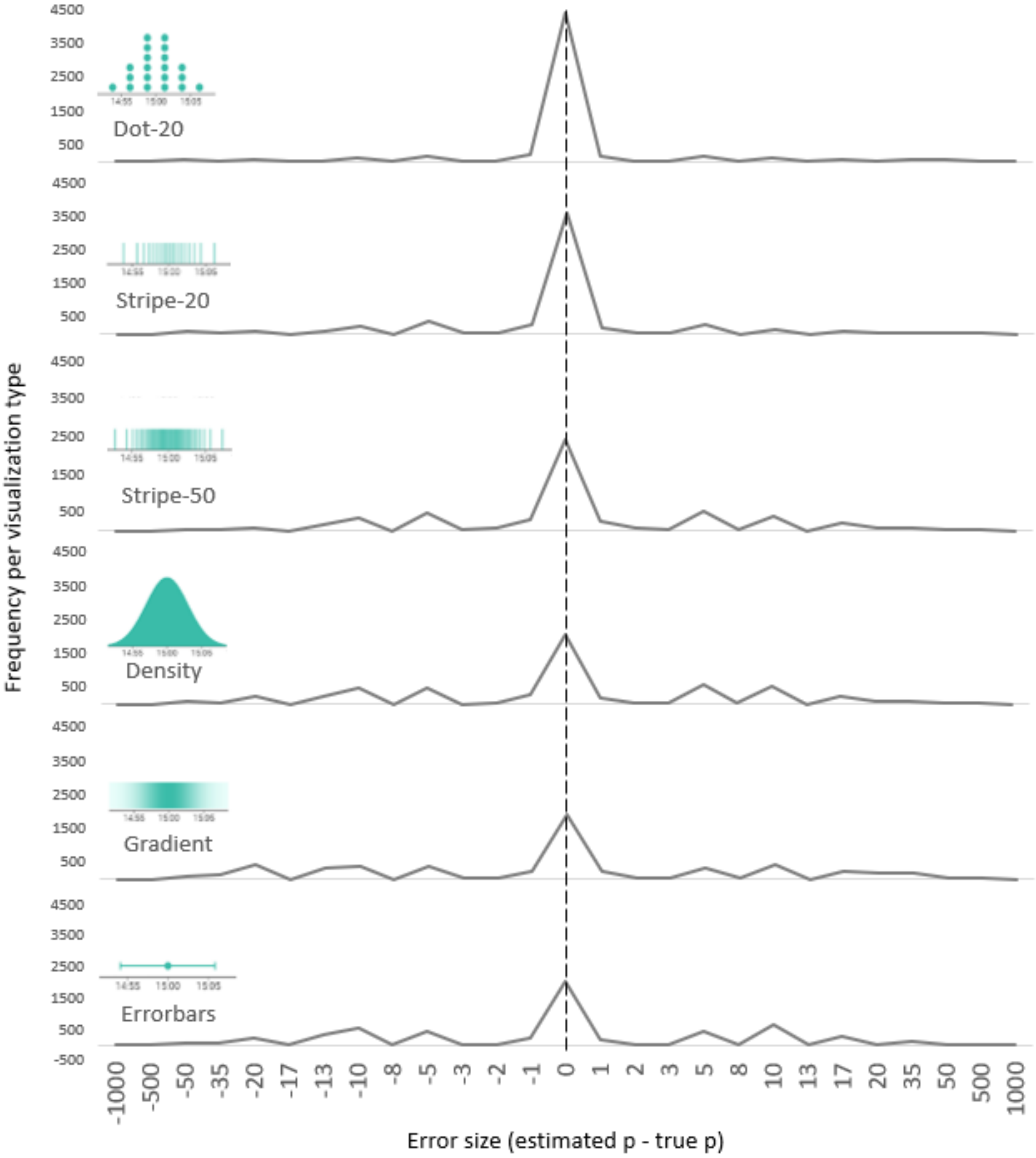
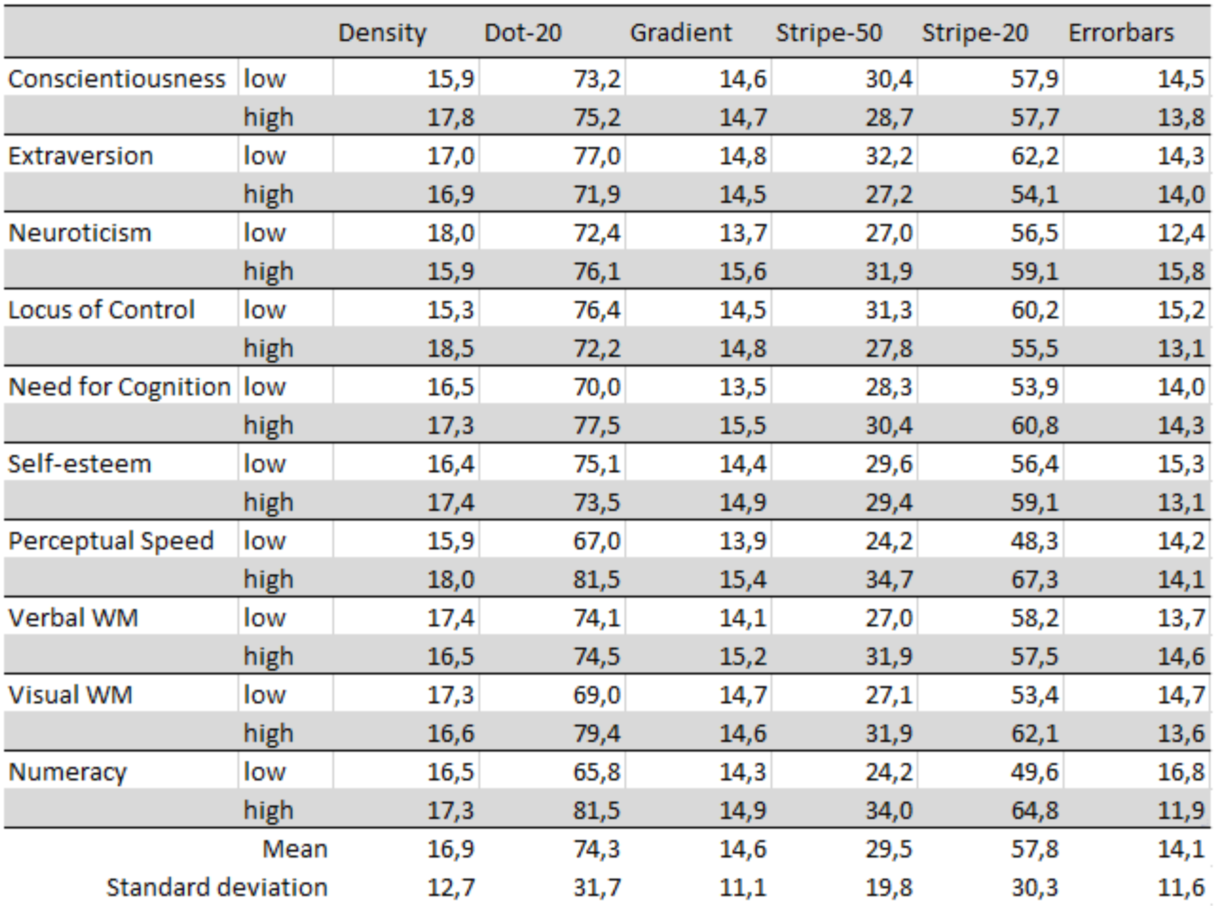


Figure 16. Accuracy curve before log transformation. Note: bin sizes are irregular.

Appendix D: Median split

Table 13.

Average percentages correct on the graph task (Q2 and Q3) per high/low group after performing a median split for every user characteristic.



		Density	Dot-20	Gradient	Stripe-50	Stripe-20	Errorbars
Conscientiousness	low	15,9	73,2	14,6	30,4	57,9	14,5
	high	17,8	75,2	14,7	28,7	57,7	13,8
Extraversion	low	17,0	77,0	14,8	32,2	62,2	14,3
	high	16,9	71,9	14,5	27,2	54,1	14,0
Neuroticism	low	18,0	72,4	13,7	27,0	56,5	12,4
	high	15,9	76,1	15,6	31,9	59,1	15,8
Locus of Control	low	15,3	76,4	14,5	31,3	60,2	15,2
	high	18,5	72,2	14,8	27,8	55,5	13,1
Need for Cognition	low	16,5	70,0	13,5	28,3	53,9	14,0
	high	17,3	77,5	15,5	30,4	60,8	14,3
Self-esteem	low	16,4	75,1	14,4	29,6	56,4	15,3
	high	17,4	73,5	14,9	29,4	59,1	13,1
Perceptual Speed	low	15,9	67,0	13,9	24,2	48,3	14,2
	high	18,0	81,5	15,4	34,7	67,3	14,1
Verbal WM	low	17,4	74,1	14,1	27,0	58,2	13,7
	high	16,5	74,5	15,2	31,9	57,5	14,6
Visual WM	low	17,3	69,0	14,7	27,1	53,4	14,7
	high	16,6	79,4	14,6	31,9	62,1	13,6
Numeracy	low	16,5	65,8	14,3	24,2	49,6	16,8
	high	17,3	81,5	14,9	34,0	64,8	11,9
	Mean	16,9	74,3	14,6	29,5	57,8	14,1
	Standard deviation	12,7	31,7	11,1	19,8	30,3	11,6