Priming and Anchoring Effects in Visualization

André Calero Valdez, Martina Ziefle, *Member, IEEE* and Michael Sedlmair, *Member, IEEE*

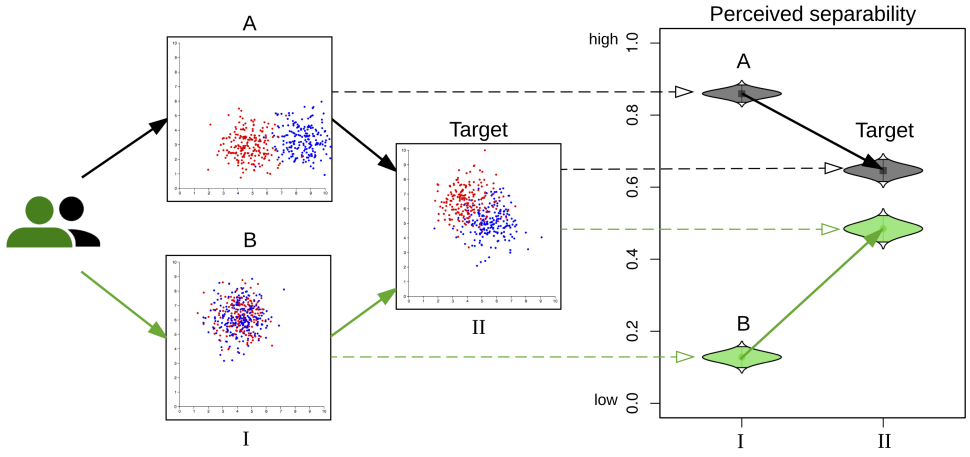

Fig. 1. The exposure to a different stimulus changes the subsequent judgment of separability in a scatterplot target. Means and 95% confidence intervals of judgments depicted as catseye plots (data from study 4 in this article).

Abstract— We investigate priming and anchoring effects on perceptual tasks in visualization. Priming or anchoring effects depict the phenomena that a stimulus might influence subsequent human judgments on a perceptual level, or on a cognitive level by providing a frame of reference. Using visual class separability in scatterplots as an example task, we performed a set of five studies to investigate the potential existence of priming and anchoring effects. Our findings show that—under certain circumstances—such effects indeed exist. In other words, humans judge class separability of the same scatterplot differently depending on the scatterplot(s) they have seen before. These findings inform future work on better understanding and more accurately modeling human perception of visual patterns.

Index Terms—Perception, Anchoring, Bias, Scatterplots, Visualization, MTurk Study.

1 Introduction

Visual perception is the main epistemic source of information for understanding our environment. One assumption of visualization is the consistency of visual perception. Given the same stimulus and the same person, one should 'see' the same thing every time. However, visual perception is both a top-down and bottom-up process or, as phrased by Kinchla, a 'middle-out' process [30]. Both higher- and lower-order aspects of visualization affect perception, and prior stimuli might bias perception and change what is seen.

Human biases play a major role in psychology research. The goal is to understand the underlying mechanisms that shape 'irrational' behavior and if necessary provide means to counteract these biases. Many biases are shaped by heuristic cognitive processes and their inherent flaws [60]. Humans, for instance, are known to overestimate the risk of becoming a victim of violent crime if they have been exposed to violent shows on television [48]. This mechanism is caused by the availability heuristic: 'because I remember crime easily, it must be prevalent often'. Similar effects are present in other judgments and can be attributed to the effort to handle noisy input signals [26]. Since visualizations are increasingly used in decision making, it is necessary to understand how cognitive biases might distort decisions from visualizations.

- *André Calero Valdez is with RWTH Aachen University. E-mail: calero-valdez@comm.rwth-aachen.de.*
- *Martina Ziefle is with RWTH Aachen University. E-mail: ziefle@comm.rwth-aachen.de.*
- *Michael Sedlmair is with the University of Vienna. E-mail: michael.sedlmair@univie.ac.at.*

In this work, we focus on priming and anchoring effects, and the question of how far such effects might play a role in visualization. Priming effects describe phenomena in which human responses are influenced by a preceding perceptual stimulus [36]. For instance, participants who are asked to complete the word 'so_p' are more likely to pick the word soap, when they see a picture of a shower before. In contrast, participants, who see a picture of bread and butter, are more likely to fill it with 'soup'. Anchoring effects, on the other hand, describe the phenomenon that a previous stimulus provides a frame of reference, that is, an anchor. This anchor aids judgments, even if the stimulus is completely unrelated. For example, asking someone the amount of calories of a carrot, anchors later judgment on the amount of calories of ice-cream [40]. People then underestimate the latter. Or even more odd, telling someone the birth year of Mark Twain can affect their estimation of the length of the Mississippi [42, 65]. Priming is based on neural pre-activation, that means similar stimuli are recognized more easily, because their neural correlates are already 'warmed up'. Anchoring is (presumably) based on priming and describes the mere effect that judgments might be biased, or 'anchored', towards a preceding stimulus. Based on this line of research in psychology, we hypothesized that similar priming and anchoring effects might be at play in 'low-level'[1] perceptual judgments of visual encodings, such as the judgment of class separability in scatterplots [52]. As a concrete example, we were wondering whether an ambiguous scatterplot as shown in Fig. 1(Target) might be judged differently after seeing a stimulus with clearly separable classes, as in Fig. 1(A), as compared to a stimu-

[1] One must note that low-level is here judged from a computer science perspective. A neuroscientist or perception-psychologist would rate this as high-level.

lus with clearly non-separable classes, as in Fig. 1(B). In case priming effects in visualization really exist, we believe that their impact on visualization research can be large. Current visual perception models of patterns such as correlation [21, 29, 46], class separability [5, 54], or cluster separation [43] assume that human visual judgments are more or less consistent, and do not take potential biasing effects as the priming effect into account. In tasks where standardized responses are critical, controlling for such effects can help reduce variability.

In this work, we set out to make some first steps towards an understanding of priming and anchoring effects in visualization. To do so, we contribute a series of five Mechanical Turk and lab studies with overall 726 participants, using class separability in scatterplots as an example task.

Our findings show that both priming and anchoring effects are present in the repeated judgment of separability in scatterplots. The effects become detectable at different experimental setups. Anchoring effects are seen for few repetitions in the experiments, while priming effects were seen for larger numbers of repetitions. In repeated exposures, separability judgment was distorted at about 7% from the previous stimulus.

2 Related Work

Perception has always been a major driver in visualization design and research. Finding ways to represent data that helps end-users to draw correct conclusions from visualizations can only be done when we understand how people see. We need to understand how different representations translate to perception, and how perception then translates to mental representations.

Our goal is to study biases in low-level visual judgments, using the case of class separability in scatterplots as an example. To contextualize our work, we first review how perception of such tasks is currently modeled in visualization. We then provide a brief review of the relevant background literature in psychology on priming and anchoring effects.

As a full review of the literature is beyond the scope of this article, we would refer the interested reader to the works of Tversky and Kahneman [60] for biases, the work of Gescheider [17] for priming and methodological aspects, or for a more easy read to Kahneman's 'Thinking, fast and slow' [28].

2.1 Modeling Visual Perception

Plenty of work on modeling the perception of different low-level visual patterns exists. Rensink and Baldrige [45], for example, measured how *correlation* can be estimated from scatterplots, using adaptive processes to detect just-noticeable differences (JND) between scatterplots. Furthermore, Rensink found that judgments of scatterplots are less influenced by the shape of the dot cloud, but indeed more by the shape of the underlying probability distribution [47].

Using crowd-sourcing, Harrison et al. [21] matched nine visualization types of correlation data to their judgments, finding that the just-noticeable difference can be modeled using a Weber-Fechner Law. This means, for example, that two scatterplots can be judged as different more easily for more pronounced levels of correlation. The difference from $r = 0.9$ to $r = 0.85$ is more easily detected than the numerically same difference of $r = 0.1$ to $r = 0.05$. However, this depends on the direction of correlation and the type of visualization. Building on that work, Kay and Heer [29] argue that this model, however, does not include individual differences of perception. Using Bayesian estimation, they were able to refine the model to include such individual differences. However, they also find that scatterplots show very little inter-individual differences, that is, low variance and high precision.

Another low-level task that has gained considerable attention in the visualization literature is visual class separability in scatterplots. This is also the task we are focusing on. Conducting a qualitative analysis of 816 scatterplots, Sedlmair et al. have characterized different visual factors that are at play in the perceptual separability of color-coded classes in scatterplots [52]. Building up on this work, Aupetit and Sedlmair [5] modeled the human perception of class separability with different neighborhood graphs and purity functions [5]. Their approach, however, is based on modeling only clear-cut cases, that is, scatterplots that are either labeled clearly separable or non-separable by humans [50]. Borderline cases were excluded in the design and evaluation for simplicity reasons. Our work focuses exactly on these borderline cases, with the goal to better understand how reliable human judgments are for those, and in how far they are prone to perceptual biases.

Much other work exists that focuses on aspects that relate to models of visual perception. Lewis et al [32], for instance, looked at more coarse-grained differences between expert and novice judgments in class separability tasks. They found that while humans often disagree with automatic quality measures, class separation seems to be a task, that can to a large extent be solved by novices. This finding supports, for instance, our choice of using crowd-sourcing to conduct some of our studies. Others have investigated the limits to visualization. Haroz and Whitney [20] looked at how attention capacity limits visualization. Micallef et al. investigated the limits of humans in terms of Bayesian reasoning abilities [37], while Armstrong and Wattenberg looked at potentially incorrect conclusions that humans might draw from mixing effects [3]. All these papers share with us the idea of investigating specific aspects of the human perception and cognition that are relevant to visualization design and usage.

Having good perceptual models of the human visual perception can be useful in different ways. Generally speaking, they either can be used to infer which visualization is most helpful given a certain set of data, or they can automate the process of setting parameters of a specific form of visual encoding. Many examples exist, such as finding good parameters for scatterplot visualizations [38], aspect ratios [58], multi-table views [33], and other types of visualizations [64]. Another viable line of work is using such perceptual models to find interesting projections of high-dimensional data that reveal perceptually relevant aspects in data. One example for this approach are the scagnostics measures that model human perception of certain visual patterns in scatterplots [63], but many others exist [5, 6, 43, 54]. Sacha et al. propose to compare these perceptual measures with measures on the actual data [49]. If both measures agree, it is a good sign that the patterns 'really' exist (or not). Along similar lines, Kindlman and Scheidegger [31] provide an algebraic model to understand which transformations are invertible and which ones follow invariates. Ideally, we can assure that changes in data are also reflected by changes in a resulting visualization that humans perceive.

2.2 Biases in Visual Perception

While modeling perception is common in visualization, to the best of our knowledge misjudgment has not yet been addressed in repeated usage of such visualizations. Typically, repeated-measures designs, or stair-case procedures [17] are used to reduce the effect of ordering on judgment. Nevertheless, systematic effects, or biases, might exist.

In order to understand biases, one must look at the long history of research on such effects in psychology and visual perception. One of the early researchers interested in how perception is affected by prior experiences or context is Harry Helson [24]. In his 1947 APA article [23], he showed how different intensities of perception, such as weight or illumination, are evaluated against different levels of perception. Between two different judgments (heavy vs. light, bright vs. dark) there is a neutral zone, the adaptation level. The location of this zone is, however, affected by context. Stimuli may 'anchor' perception and thus bias judgment. This type of bias can be assumed to exist purely on a perceptual level. On higher levels of cognition, other biases exist that may also affect judgments.

Tversky and Kahnemann famously investigated decision making under uncertainty [60]. They used biases to understand the heuristic decision-making processes underlying these biased judgments. One famous cognitive bias is caused by **anchoring**. Anchoring describes the phenomenon that a given stimulus affects later judgments in the direction of the previous judgment, even if both stimuli are completely unrelated. A vivid example lies in the following experiment: When asked whether the age of Gandhi at his death was higher or lower than 9 years, the average guess of his actual age at death was lower than when asked whether his age was lower or higher than 141 years [56].

Table 1. Study overview: main results and samples before and after data cleaning (e.g. speeders, outliers)

Study	Main Result	Sample	Used
Pilot 1	Unclear separability in SepMe [5] possibly evokes **priming** biases.	200	180
Pilot 2	Task: **Identified target stimuli** with high uncertainty.	47	43
Study 3	Clearly separable stimuli cause **priming** on subsequent unclear stimuli.	251	196
Study 4	A single stimulus causes **anchoring** effects on later judgments.	351	243
Study 5	In repeated exposure, every single stimulus **primes** subsequent judgment.	64^*	$64^\dagger$

^{*35,105} individual judgments, $^\dagger 28,544$ individual judgments

Although both numbers are clearly wrong[2], both influence judgment by 'anchoring'. While 'anchoring', as a result from the behavioral sciences, is descriptive in nature and explains judgment bias direction, towards the anchor, it gives no reasoning of why such an effect exists though.

Priming on the other hand is a result from the cognitive sciences [36, 59]. It refers to the phenomenon that a target stimulus is more easily recognized when it has been activated beforehand. Early models of priming consider it a memory effect [59], while modern neuroscience research follows a model of neural pre-activation [27]. Priming happens unconsciously, and is relatively independent of episodic and semantic memory. For example, priming works with amnesia patients, and under the influence of drugs. Even though it is often considered a perceptual effect, also semantic and conceptual priming exists. Priming effects are more pronounced when participants are unaware of priming [57], less motivated [34], and when primes are more extreme [25].

Both priming and anchoring are related to one another, yet crucially different. While anchoring seems to be a consequence of priming [41], it does not solely rely on the mechanisms of priming. Priming is relatively fragile, while anchoring is relatively robust [65]. This means that an anchoring occurs even when the effect is made known to the participant. The reasons seems to be that anchoring occurs because the participant generates a hypothesis at the anchor—with the help of priming—and then tests against the already anchored hypothesis. This self-generation effect [41] also leads to the underestimation of the effect of an anchor on judgments by the participants.

In our work, we are interested in studying both anchoring and priming effects more thoroughly from a visualization design and research perspective. Specifically, we take an application-driven view and study their potential presence and impact using the example of scatterplots and the task of class separability.

3 Overview

In order to investigate the presence of priming and anchoring in class separation tasks, we conducted five experiments. Each of the experiments serves a different purpose in identifying whether and how a stimulus affects judgment in later decisions. The first two studies provide the overall hypothesis and the stimuli for the later studies; study 3 and 4 try to separate the two possible bias effects and study 5 measures the strength of the priming effect (see also Table 1). Studies 1, 3, and 4 were conducted with Amazon Turk. Further details of the studies are depicted in the following sections.

3.1 Focus and Justification

Classical experiments in finding *priming effects* utilize subliminal activation (stimulus onset asynchrony $< 100ms$) [61]. This ensures that no conscious evaluation affects the judgment. In contrast, classical experiments in finding *anchoring effects* explicitly make the anchor visible and might even tell the participant to ignore it. These types of setups focus on carefully separating the effects from each other.

On an application level, however, effects are not isolated and might even have compounding effects. What we are interested in is the *amount of error* introduced by such effects in a task that is relevant and at least structurally similar to a real world task. Our goal with this application-driven choice is to increase the ecological validity, while we necessarily need to accept trade-offs in terms of precision [35].

For our studies, we chose class separability of scatterplots as an example task. Class separability provides enough uncertainty, so humans have to rely on heuristic judgments and cannot calculate the optimal solution. It also has been shown to be a task where humans are quite capable [18, 32], so we do not measure random effects. These two conditions allow measuring biases and provide a meaningful field of application. We also have worked on this task in our previous work, during which we first hypothesized about the potential existence of anchoring and priming effects [5, 50–52].

[2]He died at the age of 78.

3.2 Five Studies and Main Results

The five experiments were conducted between summer 2016 and spring 2017. The first experiment started from a design that was very close to application level, with real world data and several judgments. It was a starting point to investigating priming effects in cluster separability tasks. As effects were unclear, we continued to control variables more strictly in order to isolate the effect. The second study served to select more appropriate stimuli. We found that the distance of centroids is a good predictor of separability judgment, and based on this, picked three stimuli at both ends and the middle of the scale. Study 3 was intended to measure the effect of priming using the previously selected stimuli in a setup with very few stimuli. This setup was selected to reduce the error due to repeated exposure. Since anchoring effects might occur from training tasks, we removed the two training tasks for study 4. This study then found that a single judgment of separability works as an anchor to following judgments. Whether this was due to priming could not be determined. To answer this question, we designed a fifth and last study with a large number of tasks per participant to measure the effects of priming and anchoring in repeated experiments. We found such effects and found that they influenced judgment at about 7% of the strength of the current stimulus.

3.3 Experimental Procedures

Setup All experiments were conducted using web-based surveys (i.e., SurveyMonkey and Limesurvey) or self-coded web tools (PHP, JavaScript). Participants of studies 1, 3, and 4 were recruited using Amazon's Mechanical Turk. Unfortunately, because Turkers worked more slowly than expected, they were effectively paid 2.50$/hour, which is less than we intended (we encourage future studies to use national minimum wages as a guideline). The amount of payment is an experimental factor chosen at a very low level. To not vary this influence, we did not change the amount of payment during the studies. However, due to the ethical implications of the cumulative effect of such procedures we further used the MTurk Bonus system to increase payment after the studies to $7.25. Participants of studies 2 and 5 were recruited from students and employees of our local universities.

Data Cleaning Since we partially crowd-sourced the experiments, it is necessary to remove speeders and participants that did not take the studies seriously or misunderstood the task description. Afterwards, we should gain reliable data [22]. For this purpose we removed participants that matched the following criteria: Participants who (1) chose only one level of separability across all trials; (2) reverse coded the results, that is, high separability with centroid distances of 0; (3) coded clearly different stimuli equally; or (4) took less than 100ms per task.

Statistical Analysis We take an exploratory approach and analyze the resulting data based on effect sizes, with 95% confidence intervals as recommended by APA [2]. Note, that we specifically refrain from using null-hypothesis significance testing (NHST), which has shown to yield problems, such as over-interpretation of p-values [13, 53]. For studies 1–4, we report means of judgments as point-estimates with 95%CI.

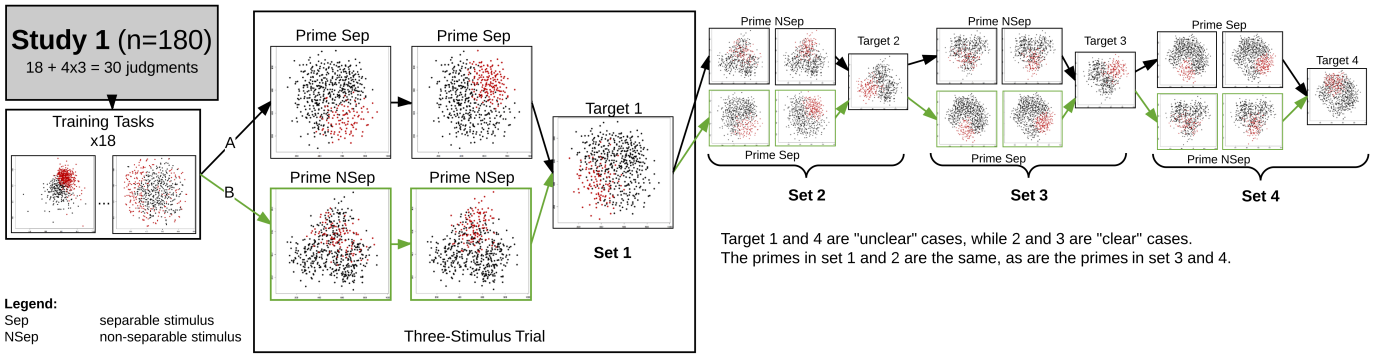

Fig. 2. Method overview of pilot study 1. Participants did either go into condition A (black) or B (green), conducting 18+12 judgments. Priming stimuli from the three-stimulus-trial sets 1 and 2 were the same, but switched between groups—same for set 3 and set 4.

Different ways of interpreting them are discussed by Cumming [13]. Study 5 differs from this approach in that we use a regression model approach. Specifically, we seek to predict human class separability based on the distance between two class centroids. We then analyze how much impact the previously seen stimulus has on the accuracy of this prediction.

3.4 Hypothesis

We generally hypothesized that priming effects and thus lastly also anchoring effects are present in repeated exposure to visualizations, similar to Mochon and Frederick [40]. Seeing a certain set of data visualized, should influence judgments on a later set of data visualized. Typically this is not desired.

4 Five Studies

In the following, we describe the five studies that we conducted in chronological order. For each study, we briefly outline necessary details of the experimental design, and discuss the main results.

4.1 Pilot Study 1 – Measuring Bias in the SepMe Dataset

The initial idea for studying biasing effects in scatterplots arose when looking at stimuli from the SepMe dataset that was used in several previous studies [50–52]. This data contains 816 scatterplots, with overall 5,460 classes. Each of these classes comes with a rating of two expert judges on a 5-point scale from clearly non-separable to clearly separable. After looking at many these plots, the intuition arose that judgments of 'unclear' cases depended to some extent on the decision made before. The initial hypothesis was, that each stimulus, that is, each scatterplot, biased the subsequent judgment as it primes the decision in one direction. However, it was unclear to us, to what extent, in which direction, and by which means.

Experimental Design To understand this effect, we designed an experiment that would measure the effect of priming in consecutive judgments. In order to set the stage for this effect, we asked participants to first judge 18 training stimuli in regard to separability. We carefully handpicked training stimuli that showed clear congruence in the existing expert judgments. So they were either clearly separable or clearly non-separable according to both judges. For training, these 18 stimuli were presented in an order that switched between clearly separable and non-separable stimuli in each turn. Since another bias in range judgments is known from frequency range theory [44], we used this procedure to establish a baseline frequency for all possible judgments. After these 18 training stimuli, we presented four sets of three stimuli (see Fig. 2), where the last stimulus was the target stimulus, while the first two were priming stimuli. Sets 1 and 4 should measure the priming effect, while sets 2 and 3 should control the non-existence of the effect in clear-cut cases. All priming stimuli, as well as target stimuli 2 and 3 were rated congruently by the expert coders. In contrast, target 1 and target 4 had different ratings from the expert coders, which were unclear stimuli. In order to prevent effects of ordering we randomized the sample in two conditions (A=black and B=green) and changed the order of the 4 tasks. As in the original study by Sedlmair et al. [51], separability was measured on a 5-point Likert scale (*clearly separable*, *separable*, *unsure*, *non-separable*, *clearly non-separable*). 200 participants took part in the Amazon Turk study, 180 remained after data cleaning.

Our hypothesis for this study was that seeing two consecutive stimuli of the same 'type' (separable or non-separable) will affect the judgment of the consequent 'unclear' stimulus. However, it should only apply to unclear stimuli (target 1 and 4). No effect was expected for clearly rated stimuli (target 2 and 3). This setup was most similar to our process of browsing these data manually before.

Results We could find some priming effects for the targets 1 and 4 (see Fig. 3). However, the results were not as clear as we hoped for. While target 1 did show a small difference in means, no clear difference was found for target 4. There is a hint of difference for this target, however, such differences also occurred during the training trials for early stimuli. As expected, no effect occurred for target 2 and target 3. Participants took about 8 minutes for the task.

From this experiment we have learned that some type of 'priming' effect might exist (see Fig. 3, target 1), However, changing both priming and target stimulus for the repetition in the same condition could have lead to a reduced effect or no effect at all (see Fig. 3, target 4). So we have conflicting evidence in the same experiment. Target 1 shows a small priming effect, while target 4 hardly shows an effect.

We also do not know how the 18 training tasks might have affected the judgment process overall. In this study we chose two priming stimuli and thus also included repetition priming [27] into the equation, possibly overemphasizing the effect. From these results we decided to generate new stimuli that removed some of the variance introduced by the different shapes of stimuli chosen for the individual conditions.

4.2 Pilot Study 2 – Finding Target Stimuli

The stimulus set in study 1 consisted of stimuli of varying degree in density, dot counts, clumpiness, shape, etc. [52]. This allowed for too many uncontrolled variables in our design, possibly affecting judgment and masking conclusive effects. As we do not expect to see very large effects, we try to avoid this by finding more homogeneous stimuli.

To better control for such confounding factors that might mask effects, we decided to reduce the task to separating a simple set of two almost identical point clouds. We thus opted for only varying a single variable, class distance, as the single most important factor indicated in Sedlmair et al.'s taxonomy of class separation factors [52].

The idea was now to generate stimuli procedurally and select stimuli that were in line with the experimental design of study 1. The priming stimuli should have congruent judgments; the target should be unclear and in the adaptation-zone. The sole purpose of the second study was to identify three stimuli that follow these requirements and that can be used in the following studies.

Experimental Design As in study 1, each stimulus contained two different clusters. Each of these point clusters was generated

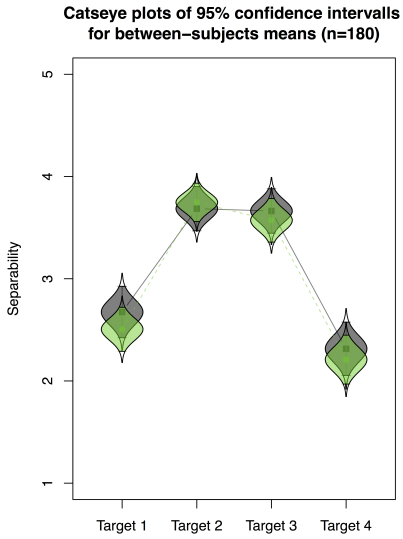

Fig. 3. Results of pilot study 1: Catseye plots show 95% confidence intervals (CIs) for the mean ratings of the two conditions, only for the targets. Target 1 shows a difference between conditions and target 4 shows hints of differing judgments between conditions.

from a bi-variate distribution ($SD_x = 1$, $SD_y = 1$) using 200 samples, a choice we made based on our previous experience working with class separability in scatterplots [5, 50–52]. Cluster distance is calculated as the euclidean distance of the centroids of the clouds. Distances were randomly chosen from the domain $[0;4]$ standard deviations. As a sampling method for normally distributed random data, we used the *Box-Muller-Transformation* (BMT). BMT is a fast effective algorithm for generating Gaussian distributions [7]. Since it was used in JavaScript and speed was an issue, this algorithm was chosen. Coloring was assigned randomly to create two groups (red/blue and blue/red) for the same cluster data.

Participants conducted 2 training tasks, before we asked them to judge 50 trials of randomly generated cluster data (see. Fig. 4). Each participant saw the same set of random stimuli in the same order as stimuli were generated using a fixed random-seed. The study was conducted with 47 students from the field of Digital Media Communication and Computer Science, 43 remained after data cleaning. 5-point separability judgments could be submitted by keyboard, allowing individual judgments to be performed relatively quickly. Participants took on average 2.5 minutes for the task.

Results To identify three representative and good stimuli, we analyzed the stimuli by looking at the variance and mean of judgments, as well as the histograms of judgments. As priming stimuli, we picked one that showed very clear separability (Sep, distance = 2.92 SD), and one which was rated as very clearly non-separable (NSep, distance = 0 SD). As a target stimulus, we picked one with a medium rating and a large variance, indicating (uncontrolled) disagreements between participants (Target, distance = 2 SD). All three stimuli can be seen in the Fig. 4 and their histograms below.

4.3 Study 3 – Measuring Priming Effects

After finding good stimuli in study 2, we sought to better isolate the possible priming effect with a very short and clear study procedure. Participants were only shown four stimuli: Sep then Target (T), and NSep then T again (reverse order for the other group). Our hypothesis was that after judging the separability of a single point cloud that is very clearly separable (Sep), the judgment of an unclear stimulus (T) is biased towards the previous judgment, namely separability. Analogously, after seeing a clearly non-separable point cloud (NSep), we expected the preceding judgment of T to be primed towards non-separability. In doing so, we sought to reduce the impact of the (previously 18) training stimuli on the target judgment.

Experimental Design In order to control for ordering effects, all participants were randomly assigned to one of two groups either starting in the non-separable (here A) or separable (here B) condition (see Fig 5). To reduce the influence of the first 'priming' stimulus on the second trial, we also needed to introduce a masking task. In this task, participants had to judge five random network graphs in regard to 'attractiveness' and 'complexity'. This task was designed on purpose to take the largest amount of time of the whole study. The task took on average 2 minutes, vs. 15 seconds experimental task.

In contrast to the previous studies, we asked participants to rate the separability of the target using a slider with no tickmarks or reference numbers. As we only measure four judgments per participant in this study, a 5-point Likert scale would have masked any subtle within-subject effects. Given that we only expect small effects, everybody would have likely defaulted to the same value, for instance '3', again. Instead, we were interested in seeing whether one single person would judge the same stimulus differently after two different 'priming' stimuli.

This study was conducted using Amazon's Mechanical turk ($n = 251$). After removing speeders as described in section 3.3 a set of 196 participants remained in the sample.

Results Study 3 yielded two interesting results. First, we did see a difference in means in the second target between subjects, as this is visible by comparing the green and black fisheye at target 2 in Fig. 6 (left side). So being in a different condition did affect judgment to a certain extent. However, no such difference was present for target 1. The means are in fact very close to each other. The group depicted in black that started with a non-separable stimulus also showed a within-subject difference between the identical target 1 and target 2. The other group depicted in green, however, showed no such difference (see horizontal lines in Fig. 6, left side). We concluded that the training task, which was one clearly separable and one clearly non-separable stimulus in that order, likely already had affected the initial judgment.

This finding becomes obvious when looking at stimulus 1 from the black group. The variance is lower than for the green group, who saw this stimulus after the masking task (see Fig. 6, stimulus 2, green catseye plot). This could be, because the black group saw a very similar stimulus as a training task right before judging this stimulus. The same effect also shows up in the differences between the separable stimuli between both groups. The green group shows a bigger variance in their judgment at stimulus 1 than the black group at stimulus 2. Yet, the green group just saw an example of a clearly separable stimulus in the training task, which showed point clouds even further away than the stimulus shown during the experiment. This could mean that the effect of the training was present in both groups, but in different directions. It moved both non-target stimuli to the center for the green group and moved both non-target stimuli to more extreme judgments for the black group. While some sort of effect seemed to be present, the very strong agreement of both groups at Target 1, was surprising.

From these results, we hypothesize that some sort of priming seems to be present, however our training tasks could have themselves affected the judgment of the stimuli. This hinders clear conclusions about priming effects. Instead, we can say that an anchoring effect might have been at work here as well. The training tasks anchor the extrema of the judgments thus affecting judgment at trial 1, increasing the certainty for the black group and lowering the judgment for the green group.

4.4 Study 4 – Removing All Training Tasks

In order to also remove the effect from the two training tasks, we conducted a fourth study.

Experimental Design Study 4, which is also shown in Fig. 5), is almost identical to study 3. It only differed by skipping the two initial forced-choice training tasks, as we could not ensure that these two training tasks would act as 'priming stimuli' for our targets. This study was also conducted using Amazon MTurk ($n = 351$). After removing speeders, a set of 243 participants remained in the sample.

Results Fig. 6 (right side) shows the results of study 4. As study 4 is very similar to study 3, the interpretation of results is also very similar. However, the actual results in this setting are different. The

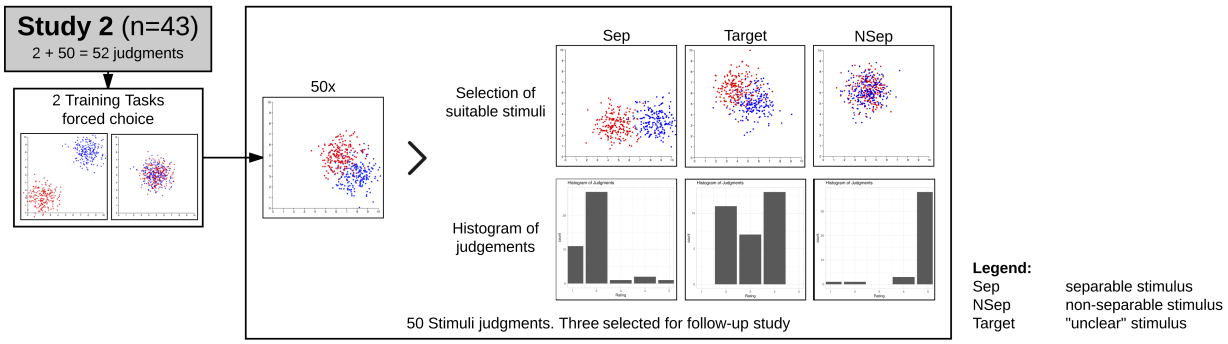

Fig. 4. Method overview of study 2: Participants were asked to rate the separability of 50 stimuli on a scale of 1 to 5. Thee three most fitting stimuli are depicted on the right.

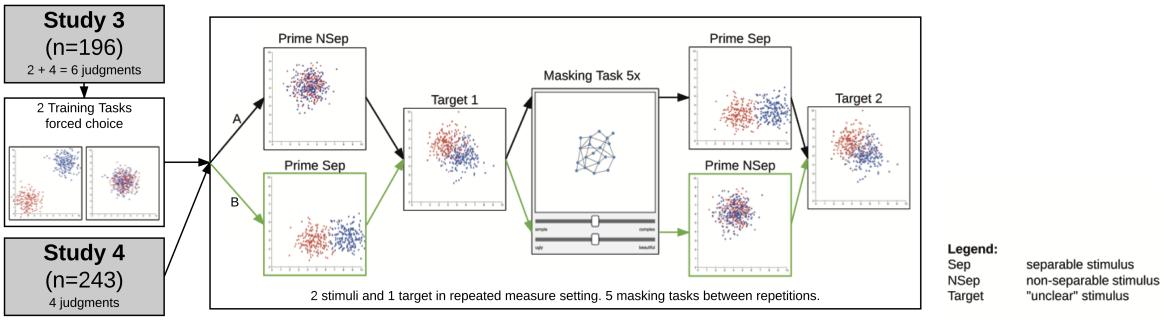

Fig. 5. Method overview of study 3 and 4. The studies differed in either having or not having two training tasks. The masking task was chosen, as it was similar (look at points, set sliders), but could not be a prime (only one cluster).

within-subject effect of the black group disappeared for both targets. In contrast, the between-subjects effects are much more pronounced at all stimuli-locations. The green group clearly rates separability of all stimuli lower than the black group. Interestingly, judgments of the same non-target stimuli also show differences between both groups. For example, the stimulus 1 of the green group is judged as less separable than stimulus 2 of the black group (same stimulus). It seems to be the case that the first stimulus anchors judgment so strongly in this setting, that no priming affects the judgment of target 2. Another explanation could be that the participants remember the target stimulus and pick a similar value as before.

We conclude from this experiment that anchoring effects are present in separability judgments and do indeed affect judgment of later scatterplots. This might nevertheless have been caused by the short experimental frame of only 4 judgments. The participants could not have established a consistent frame of reference as predicted by frequency range theory [44] and adaptation level theory [23]. Therefore a frame of reference is chosen at the first stimulus and then mapped to the consequent ones. This however would mean, that priming effects in a narrower sense would not be at work, as no within-subject effect can be seen.

4.5 Study 5 – Prediction of Rating Judgment

Since both studies 3 and 4 yielded partly conflicting results, we decided to conduct a fifth study. Study 3 indicated a priming effect exists, as the same participants (in this case the black ones) rated the same stimulus differently depending on the priming stimulus. However, it was inconclusive, as the other group showed no such effect (the green ones). There should be no anchoring effect detectable between groups, as both groups saw the same training tasks to "calibrate" their judgment. Study 4 showed no such difference (i.e., no priming effect), but a strong difference between the groups. This difference can be interpreted as an anchoring effect, as the strong difference in the first stimulus acts as an anchor for rating all following stimuli in respect to this first stimulus. Since there is no training stimulus, the rating task is now "calibrated" with respect to the first stimulus.

In order to find out whether within-subject effects, and therefore priming effects, do still occur, we decided to conduct a study that was in stark contrast to study 3 and 4. While studies 3 and 4 relied on many subjects and few judgments, study 5 does the opposite. The purpose was to detect the effect of priming in a long series of judgments. In this setting we can go back to 5-point ordinal judgments, as multiple judgments are made for each individual. These are quicker and facilitate the "many judgments per participant" design.

The hypothesis is that priming effects do occur in long term usage of visualization. We measure this by regressing the judgments of participants. Our goal of this regression is to predict the judgment of participants from the distance of the two class centroids. Large distances should lead to separable judgments, low distances to non-separable judgments. We now can build two models, one which only uses class distances, and one that uses current class distances **and** the previous class distance. If the latter one explains more variance, we can conclude that priming effects are present. A method to compare two models with different parameter counts (i.e., one parameter for no priming effects, two parameters for priming effects) is the AIC (Aikike Information Criterion) [1]. This criterion evaluates the increase of explained variance or more precisely the decrease of residuals against the additional use of parameters[3].

Experimental Design The setup of this study was derived from study 2 by extending it to maximum 1,000 trials. Participants were asked to judge separability of randomly generated point clouds (see Fig. 4). Only this time, the study would not stop after fifty trials, and was conducted with researchers and students with experience in visualization or HCI: eleven colleagues from the authors, and 53 HCI students from University of Vienna, all unfamiliar with the experimental design. No rewards were given to colleagues. Students could select

[3]One can always increase the explained variance, by adding more parameters to a model. The AIC aims to control this by penalizing against more parameters.

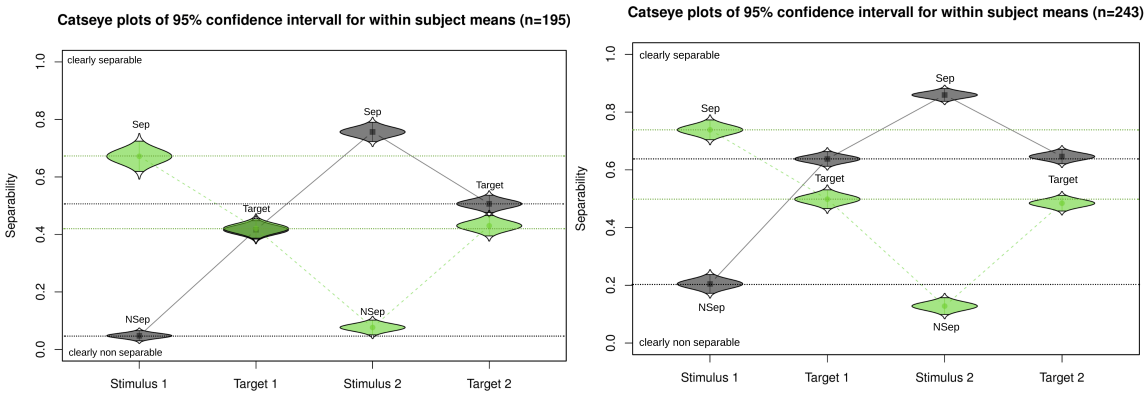

Fig. 6. Left: Catseye plots of mean judgments of study 3 (n=196). A within-subject effect can be seen for the black group, and a between subject effect at Target 2. Right: Catseye plots of mean judgments of study 4 (n=243). No within-subject effect can be seen for either group, yet a between subject effect is present at both targets.

this task as a voluntary homework and gather extra points for class. We asked 64 coders to judge as many randomly generated plots as their concentration would allow. The coders were asked to take a break every 15 minutes and only judge as long as they felt they were not making mistakes.

Results The 64 coders rated a total of 35,105 visualizations (7,257 removed as outliers). All first (one for each coder) 'non-primed' stimuli were then dropped. Only 10% of all judgments were rated as 'unclear' (see Fig. 7). Most judgments were rated as clearly non-separable (n=8,634).

In order to understand the effect of priming in so many judgments we applied regression modeling to the judgments. We used both linear regression and ordinal logistic regression to compare both models. Since judgments were done by key-presses (from 1 to 5), ordinal regression is the more accurate fit, but linear regression is easier to understand.

For linear regression models we report coefficient estimates with standard errors, t values and probabilities of parameter estimates. We report adjusted r^2, degrees of freedom and Anova results (F-Measure, p-value) to compare the model against a simple mean-model. For the ordinal logistics regression, we provide coefficient estimates, standard errors and t-values for coefficients as well as intercept estimates standard errors and t-values for individual level changes. To provide a more meaningful interpretation we also provide odds-ratios for level changes including 95% confidence intervals.

A multiple **linear regression** model using both centroid distances, the current and the previous stimulus or simply *CDistance* and *PDistance*, yielded a good model fit ($F(2,27845) = 22,510$, $p < 2.2e^{-16}$). The residual standard error was 0.8943, multiple $r^2 = 0.6179$ and adjusted $r^2 = 0.6178$. The strongest predictor was, as expected, the current centroid distance. The intercept was near 1, as expected. The intercept refers to the judgment that is picked for a distance of zero. It should be near 1, as our judgment scale started at 1. The model then predicts a user's *rating* as follows:

$$rating = 0.613 + 0.996 \times CDistance + 0.073 \times PDistance$$

That is, each unit of increased distance, increased separability by approx. 0.996 (on the limited scale from 1 to 5, see Tab. 2). The distance of the centroids in the previous stimulus increased the separability by approx. 0.073 for each unit in difference. Both predictors improve model fit. This indicates that a priming effect exists. However, one must consider that several assumptions are violated in this approach. The error is not normally distributed, as the scale is limited, and the distance is limited by zero.

To overcome the limitations of linear regression with ordinal data, we applied **ordinal regression**. Here, we want to predict the likelihood of an ordinal judgment of a participant (1, 2, 3, 4, 5) from the distance of the centroids of the clusters in the current and previously seen stimulus.

Table 2. Coefficients of Multiple Linear Regression Model using two predictors, rounded to three decimal places. (CDistance = distance of both centroids, PDistance = Distance of centroids in previous stimulus)

	Estimate	Std. Error	t value	p value
(Intercept)	0.613	0.014	44.49	$< 2e-16$
CDistance	**0.996**	0.005	211.80	$< 2e-16$
PDistance	**0.073**	0.005	15.45	$< 2e-16$

So if we know the current clusters are 1 unit apart, and the previous ones are 3 apart, we can calculate the odds for each judgment on the ordinal scale. We again use two models. A one- and a two-predictor model to estimate the effect of the priming stimulus. The first to model no priming, the second to model priming. The resulting model with two predictors yielded a residual deviance of 59,708.84 and an AIC of 59720,84 (see Tab. 3).

This means that each increase in the current centroid distance of one unit increases the likelihood of picking a higher judgment with the odds of 7.94 (i.e., an increase of 694%). Furthermore, this means that each increase of centroid distance in the previous stimulus increases this likelihood of picking a higher judgment with the odds of 1.18 (i.e., an increase of 18%). Tab. 4 shows also the confidence intervals.

Table 3. Coefficients and Intercepts of Ordinal Regression Model using two predictors

Coefficients

	Value	Std. Error	t value
CDistance	2.072	0.016	127.20
PDistance	0.170	0.010	16.19

Intercepts

	Value	Std. Error	t value
1—2	2.602	0.035	73.438
2—3	4.641	0.044	105.301
3—4	5.439	0.048	114.349
4—5	7.764	0.058	134.166

Residual Deviance: 59,708.84, AIC: 59,720.84

According to the AI-criterion [1], a model without the previous stimulus difference as a second predictor is inferior ($AIC_{1Pred} = 59,9982.24 > AIC_{2Pred} = 59,720.84$). According to Burnham & Anderson [9], this means that the model with priming is $2.9e^{56}$ times ($= e^{((AIC_{1Pred} - AIC_{2Pred})/2)}$) more likely to explain more information than the model not assuming priming.

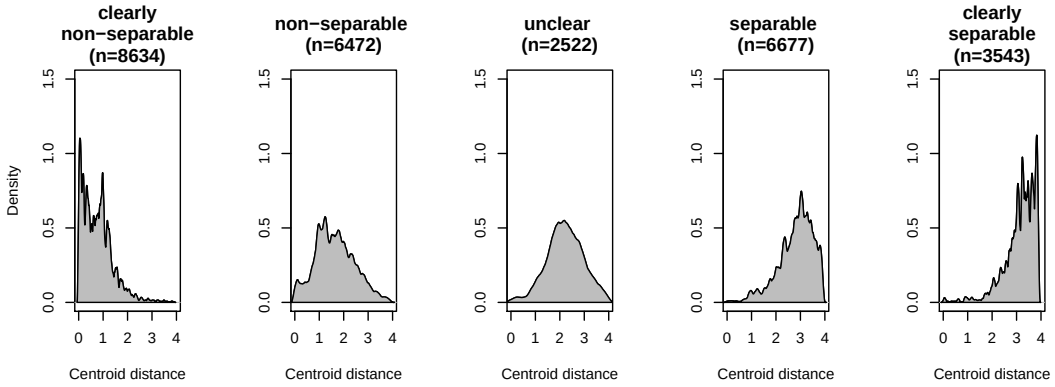

Fig. 7. Kernel density estimation using the Epanechnikov kernel of all judgments plotted by judgment category.

Table 4. Odds ratios from the observed information matrix after optimization of ordinal regression model using two predictors

	Odds Ratio	Lower CI	Upper CI
CDistance	**7.942**	7.693	8.200
PDistance	**1.185**	1.161	1.209

We conclude from these models that priming effects do occur in repeated class separability tasks. Since stimuli were generated in random order such a consistent effect of the previous stimulus can only be explained by priming. Anchoring effects would diminish over the trials, as both separable and non-separable anchors occur randomly throughout the experiment. However, a compound effect can not be totally ruled out. The strength of priming is relatively weak, as expected, but consistent and dependent on the size of the prime.

5 Discussion

In our research, we found evidence for the existence of priming and anchoring effects in class separation tasks of bi-variate normally distributed scatterplots. This means—ceteris paribus—that deciding whether two clusters are separable, does not only depend on how far they are apart, but also on how far previously seen clusters are apart. In other words, the *dynamics* of repeated exposure to a visualization does play a role in their interpretation.

In this section, we will put our findings in context to how previous research has looked at biases and related effects. We will then revisit our choice of addressing this research question from an application-driven angle, by starting from an experiment that was close to the application level, and gradually removing confounding factors. Our approach has some drawback and caveats, which we will discuss and evaluate. We also propose how to address some of these limitations and outline where new opportunities of future work can be derived.

Relation to Previous Work

Our work focuses on better understanding *temporal effects* of repeated exposure to visual stimuli, such as priming and anchoring effects. So far, the lion's share of perceptual visualization research has focused on contextual factors that are present *simultaneously*. For example, Webster and Leonard [62] looked into the effect of how the context affects the perception of the color white. They found that on a very low-level of perception the color white is adjusted to the surrounding colors. The perceptual system performs an 'auto-white balance' depending on the visual context. To our knowledge, there is much less work on how previous exposure to a visualization affects perception, and the biases that might stem from looking at sequences of visualization stimuli. Utilizing the same visual encoding over and over again is not uncommon in real-life applications. In terms of scatterplots, think for instance, of using the venerable approaches of projection pursuit [15], or grand tours [4].

Another source of interesting related effects stems from inter-individual differences. The need to model such effects is obvious, as different people might see the same thing differently. Previous work [12,55] has shown that the inclusion of inter-individual differences can add another source of systematic 'errors'. Adapting a visualization and correcting for such individual human factors (e.g., perceptual and cognitive abilities [11], personality [19], technological affinity [8, 19]) can help users make better decisions. For example, users with higher cognitive processing speed were found to be more effective at deriving facts and insights from a visualization than others [11]. Such inter-individual differences become even more important when tasks become more complex [39]. A quick glimpse into judgments in our study 5 hints at relatively large and systematic inter-individual differences. Some users prefer to pick only values at the higher end of the scale, while others pick only in the lower end of the scale, while yet others leverage the full scale. Although our investigation of priming and anchoring effects is robust against such inter-individual differences, it poses many interesting open questions for future work: Are these inter-individual differences caused by choice or perception? Do different rating 'styles' relate to different personal traits [10]? Do different users require different counter-measures to prevent misjudgment caused by priming or anchoring?

Studying Biases from an Application-driven Angle

The focus of this article lies in applying an application-driven methodology to a low-level phenomenon. With our research, we also wanted to demonstrate this methodological approach and that it can be used to study such biases from a more application-driven lens. The general methodology to detect priming effects already exists since the 1950s. So far, such effects have only been studied under highly-controlled conditions though, leaving a large gap to actual applicability and ecological validity in real-world tasks. One could ask whether millisecond differences in stimulus recognition matter for decision making using visualization. The question we aimed at is whether these effects accumulate and leave a measurable amount of error. For this purpose, our methodology tries to bridge the higher-level perspective on priming from a neuro-psychologist's view with a low-level perception view of a computer scientist. Starting form an application-driven perspective with real-world data, we increased the level of control in our experiments step by step, until we arrived at a level of abstraction that allowed us to isolate, detect, and quantify such biasing effects. Of course, this approach comes with its own caveats, as every methodological choice needs to make certain trade-offs [35]. By following a more application-driven approach, we specifically open our results to other types of interpretations. One could argue that, e.g., our effects could have been caused by the random generation of stimuli (studies 2–5). The effect could be exclusive to the stimuli we used and not depend on the regressed variables. The normally distributed error in this process, should however cancel out in longer studies, such as study 5. The effects could also have been caused by the individuals that chose longer exposure times, or the effect could have been governed by accidental mis-judgments.

Still, we believe that our interpretations do make sense in the light of theoretical explanations from the psychological models. There *should* be an effect of priming and there should be an effect of anchoring, if

the theory [36, 59, 60] is correct. The size of the effect is arguable, and needs further dissection, but this is where our research provides a starting point. The experiments conducted for this article serve as a base-line with relatively 'normal' parameters (i.e., 200 data points, 2 clusters, bi-variate normally distributed clusters). We picked these parameters from previous work [5, 50–52].

We also believe that by reporting all five studies including the first relatively inconclusive study, we allow the reader to understand how such an approach can be applied to different types of visualizations and biases. Simply reporting the last study, might be sufficient for providing evidence for the effect, but following the process shows how studies have to be adopted to separate effects and thus estimate their influence. It also helps to avoid known methodological pitfalls [13, 66].

For other researchers investigating similar effects using MTurk, we recommend carefully adjusting the payment in pilot studies and gradually extending sample sizes to ensure that task descriptions contain realistic timings and payment levels equate to task completion times. We fixed payment too early on a level that was too low, retrospectively. We reimbursed participants through MTurks bonus system after the study to ensure payment according to minimal wages in the US.

Limitations As with every empirical study, our work does not come without limitations. Our general approach was of exploratory nature, with limitations in the individual study designs, and the statistical implications of the underlying decisions we made.

Study 1, for example, is subject to variations of multiple variables in any condition, as we changed ordering and the stimulus selection. No study controlled the timing in which judgments were made. Typical studies on priming prefer subliminal presentation of primes (i.e., exposure of $< 100ms$), as this clearly separates from anchoring effects. However, such approaches were not feasible in a web-based Mechanical Turk setting and did not fit our application-driven focus. Judgment input was changed during the experiments (Likert scales for study 1, 2 and 5, sliders for study 3 and 4). Both types of inputs cause differences in exposure to the priming stimulus, affect judgment, and affect error rates. No control of presentation was performed. Differences in users displays (e.g., resolution, color representation) could have influenced the results.

Studies 3 and 4 showed bi-modal distributions of both target stimuli. We assume this was due the users' potential urge to 'take action' and not simply accept the default value, which was set to the center of the slider. In cases where the default value might be a sensible option, deciding in which direction to move the slider could lead to such a distribution. Users first decide to move the slider either left or right, and then move it a certain distance. Also, since we did not use tickmarks on the slider, users could not base their judgments on a number. If we assume that users have no inherently preferred direction, when presented with such a (semi-)forced-choice, then on average the effect should cancel out. A follow-up study with no default option and tickmarks would be interesting to shed further light into this question.

In no study did we explicitly instruct participants to judge separability spatially. In theory, we could have accidentally caused participants to judge separability because of chromostereopsis. Since we picked highly saturated blue and red, viewers could have moved the blue points into a lower viewing plane and considered this for separability. However, the results from study 5 indicate that on average spatial distance was indeed used to assess separability.

From a statistical point of view, judgments are not normally distributed, as they are limited by the judgment scale. And neither are the stimuli, whose centroid distance was also limited between 0 and 4 standard deviations. Thus a Bayesian approach to modeling could reveal more faceted understanding of effects. Our approach might overestimate the size of effect here.

However, we believe the overall results and the diversity of studies outweigh these limitations. Every experiment needs to make a careful trade-off between the many influences that add onto each other, specifically in more complex, application-driven settings.

Future Work Future work can immediately be seen in changing parameters in our experiments and extending it to other types of visualizations, tasks, and biases. Once a more consistent picture of the influence exists, one can start trying to counter-act the effect and wonder what a better model of perception means for the field of visualization. In our work, we started using 200 samples bivariate distributions merely shifting centroid locations. We can use more or less samples, other distributions, or more classes. Other metrics, such as clumpiness, density or other factors [52, 63], can be varied to investigate their effect on priming or anchoring as well. As already partially addressed in study 1, *repetition priming* could be investigated by providing a series of priming stimuli before showing the target stimulus. Instead of judgment tasks, priming effects on just-noticeable differences could be investigated as well (e.g, by stair-case procedures [16]). By changing the *instructions* of the task (e.g., 'Make sure the previous stimulus does not influence your judgment') moderatability could be investigated. Do instructions counter-act the effect, or does the visualization have to start to become 'deceptive' by distorting data? What about parallel presentation, as in scatterplot matrices or small multiples?

As mentioned before, even though our results are consistent in showing an effect, we have not yet demonstrated that such an effect can actually be counter-acted in a visualization. By predicting the effect in a study setting, we could artificially move clusters apart and see whether judgments are more accurate to the 'intended data'. This would be a first step in using the effect to improve visualizations. This idea goes in a similar direction as user-adaptive visualizations, where the visualization adapts to users capabilities, preferences and—with the findings from future research—possibly to the users biases as well. The interaction of bias and inter-individual differences is particularly interesting, as some biases are affected by inter-individual stable and unstable factors (e.g., mood or expertise [14]).

The visual encodings in our experiment were completely free from interpretation, or at least no explicit one. How would priming affect judgment if the data represented data from different fields of study? Is cancer-data more susceptible to bias than sales-data? What if each cluster refers to different nationalities and separability indicated different capabilities? Would such a setting increase anchoring effects from hypotheses? Would prejudices translate to bias in visualization? Extending this research to other types of visualizations, tasks, and biases is relevant for an adequate model of human perception. In the long-term, such research could enable visualizations to communicate what they intend to and help users to overcome their biases.

6 Conclusion

Much visualization research has aimed at modeling the perception of visual patterns. Systematic errors—or biases—in this process, however, have been largely overlooked so far.

In this work, we sought to gain a better understanding of a specific type of biases, namely priming an anchoring effects in visualization. Our findings show, that in fact these effects can be observed for the task of class separability in scatterplots. However, we see our work only as a starting point rather than a final answer. To the best of our knowledge, no one has shown that biases impact people's sequential judgments of visualizations. We envision that studying such effects will become a viable area of visualization research, and hope that our work will inspire others to study visual biasing effects for other tasks, under varying conditions, and from different angles.

Acknowledgments

We thank all study participants and the very thoughtful and helpful reviewers. Thanks to Lena Oden and Felix Heidrich for helping set up MTurk studies. This work was partly funded by the German Research Council DFG excellence cluster "Integrative Production Technology in High Wage Countries", and the FFG project 845898 (VALID).

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[2] American Psychological Association. *Publication manual of the American Psychological Association (6th edition)*. American Psychological Association Washington, 2010.

[3] Z. Armstrong and M. Wattenberg. Visualizing statistical mix effects and simpson's paradox. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 20(12):2132–2141, 2014.

[4] D. Asimov. The grand tour: a tool for viewing multidimensional data. *SIAM Journal on Scientific and Statistical Computing*, 6(1):128–143, 1985.

[5] M. Aupetit and M. Sedlmair. Sepme: 2002 new visual separation measures. In *IEEE Pacific Visualization Symposium*, pp. 1–8, 2016.

[6] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 17(12):2203–2212, 2011.

[7] G. E. Box, M. E. Muller, et al. A note on the generation of random normal deviates. *The annals of mathematical statistics*, 29(2):610–611, 1958.

[8] P. Brauner, S. Runge, M. Groten, G. Schuh, and M. Ziefle. Human factors in supply chain management. In *International Conference on Human Interface and the Management of Information*, pp. 423–432. Springer Berlin Heidelberg, 2013.

[9] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media, 2003. Page 70.

[10] A. Calero Valdez, P. Brauner, M. Ziefle, T. Wolfgang, and M. S. Kuhlen. Human factors in information visualization and decision support systems. *Mensch und Computer 2016–Workshopband*, 2016.

[11] C. Conati, G. Carenini, E. Hoque, B. Steichen, and D. Toker. Evaluating the impact of user characteristics and different layouts on an interactive visualization for decision making. In *Computer Graphics Forum*, vol. 33, pp. 371–380. Wiley Online Library, 2014.

[12] C. Conati and H. Maclaren. Exploring the role of individual differences in information visualization. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '08, pp. 199–206. ACM, New York, NY, USA, 2008. doi: 10.1145/1385569.1385602

[13] G. Cumming. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, 2012.

[14] B. Englich and K. Soder. Moody experts—how mood and expertise influence judgmental anchoring. *Judgment and Decision Making*, 4(1):41, 2009.

[15] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on computers*, 100(9):881–890, 1974.

[16] M. A. García-Pérez. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision research*, 38(12):1861–1881, 1998.

[17] G. A. Gescheider. *Psychophysics: the fundamentals*. Psychology Press, 2013.

[18] M. Gleicher, M. Correll, C. Nothelfer, and S. Franconeri. Perception of average value in multiclass scatterplots. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 19(12):2316–2325, 2013.

[19] T. M. Green and B. Fisher. Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In *Proc. IEEE Symp. on Visual Analytics Science and Technology (VAST)*, pp. 203–210. IEEE, 2010.

[20] S. Haroz and D. Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 18(12):2402–2410, 2012.

[21] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using Weber's law. *Proc. IEEE Information Visualization Symp. (InfoVis)*, 20(12):1943–1952, 2014.

[22] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. In *Proc. ACM Conf. Human Factors in Computing Systems (CHI)*, pp. 203–212, 2010.

[23] H. Helson. Adaptation-level as frame of reference for prediction of psychophysical data. *The American journal of psychology*, 60(1):1–29, 1947.

[24] H. Helson. Adaptation-level theory. 1964.

[25] P. M. Herr, S. J. Sherman, and R. H. Fazio. On the consequences of priming: Assimilation and contrast effects. *Journal of experimental social psychology*, 19(4):323–340, 1983.

[26] M. Hilbert. Toward a synthesis of cognitive biases: how noisy information processing can bias human decision making. *Psychological bulletin*, 138(2):211, 2012.

[27] T. W. James and I. Gauthier. Repetition-induced changes in bold response reflect accumulation of neural activity. *Human brain mapping*, 27(1):37–46, 2006.

[28] D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[29] M. Kay and J. Heer. Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 22(1):469–478, 2016.

[30] R. Kinchla and J. Wolfe. The order of visual processing:"top-down," "bottom-up," or "middle-out". *Perception & psychophysics*, 25(3):225–231, 1979.

[31] G. Kindlmann and C. Scheidegger. An algebraic process for visualization design. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 20(12):2181–2190, 2014.

[32] J. M. Lewis, M. Ackerman, and V. de Sa. Human cluster evaluation and formal quality measures: A comparative study. In *CogSci*, pp. 1870–1875, 2012.

[33] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 13(6), 2007.

[34] L. L. Martin, J. J. Seta, and R. A. Crelia. Assimilation and contrast as a function of people's willingness and ability to expend effort in forming an impression. *Journal of Personality and Social Psychology*, 59(1):27, 1990.

[35] E. Mcgrath. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000 (2nd ed)*. Citeseer, 1995.

[36] D. E. Meyer and R. W. Schvaneveldt. Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of experimental psychology*, 90(2):227, 1971.

[37] L. Micallef, P. Dragicevic, and J.-D. Fekete. Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 18(12):2536–2545, 2012.

[38] L. Micallef, G. Palmas, A. Oulasvirta, and T. Weinkauf. Towards perceptual optimization of the visual design of scatterplots. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 23(6):1588–1599, 2017.

[39] V. Mittelstädt, P. Brauner, M. Blum, and M. Ziefle. On the visual design of erp systems the–role of information complexity, presentation and human factors. *Procedia Manufacturing*, 3:448–455, 2015.

[40] D. Mochon and S. Frederick. Anchoring in sequential judgments. *Organizational Behavior and Human Decision Processes*, 122(1):69–79, 2013.

[41] T. Mussweiler and F. Strack. Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology*, 35(2):136–164, 1999.

[42] D. M. Oppenheimer, R. A. LeBoeuf, and N. T. Brewer. Anchors aweigh: A demonstration of cross-modality anchoring and magnitude priming. *Cognition*, 106(1):13–26, 2008.

[43] A. V. Pandey, J. Krause, C. Felix, J. Boy, and E. Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proc. ACM Conf. Human Factors in Computing Systems (CHI)*, pp. 3659–3669. ACM, 2016.

[44] A. Parducci. Category judgment: a range-frequency model. *Psychological review*, 72(6):407, 1965.

[45] R. Rensink and G. Baldridge. The perception of correlation in scatterplots. *Computer Graphics Forum (Proc. EuroVis)*, 29(3):1203–1210, 2010.

[46] R. A. Rensink. On the prospects for a science of visualization. In *Handbook of human centric visualization*, pp. 147–175. Springer, 2014.

[47] R. A. Rensink. The nature of correlation perception in scatterplots. *Psychonomic Bulletin & Review*, pp. 1–22, 2016.

[48] K. Riddle. Always on my mind: Exploring how frequent, recent, and vivid television portrayals are used in the formation of social reality judgments. *Media Psychology*, 13(2):155–179, 2010.

[49] D. Sacha, M. Sedlmair, L. Zhang, J. Lee, D. Weiskopf, S. North, and D. Keim. Human-centered machine learning through interactive visualization: Review and open challenges. In *Proc. European Symp, on Artificial Neural Networks (ESANN)*, 2016.

[50] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Computer Graphics Forum*, 34(3):201–210, 2015.

[51] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 19(12):2634–2643, 2013.

[52] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. In *Computer Graphics Forum*, vol. 31, pp. 1335–1344. Blackwell Publishing Ltd, 2012.

[53] J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

[54] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. In *Computer Graphics Forum (Proc. EuroVis)*, vol. 28, pp. 831–838, 2009.

[55] B. Steichen, G. Carenini, and C. Conati. 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, 2013.

[56] F. Strack and T. Mussweiler. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology*, 73(3):437, 1997.

[57] F. Strack, N. Schwarz, H. Bless, A. Kübler, and M. Wänke. Awareness of the influence as a determinant of assimilation versus contrast. *European journal of social psychology*, 23(1):53–62, 1993.

[58] J. Talbot, J. Gerth, and P. Hanrahan. An empirical model of slope ratio comparisons. *IEEE Trans. Vis. & Comp. Graphics (TVCG)*, 18(12):2613–2620, 2012.

[59] E. Tulving and D. L. Schacter. Priming and human memory systems. *Science*, 247(4940):301, 1990.

[60] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*, pp. 141–162. Springer, 1975.

[61] D. Vorberg, U. Mattler, A. Heinecke, T. Schmidt, and J. Schwarzbach. Different time courses for visual perception and action priming. *Proceedings of the National Academy of Sciences*, 100(10):6275–6280, 2003.

[62] M. A. Webster and D. Leonard. Adaptation and perceptual norms in color vision. *J. Opt. Soc. Am. A*, 25(11):2817–2825, Nov 2008. doi: 10.1364/JOSAA.25.002817

[63] L. Wilkinson and A. Anand. Graph-theoretic scagnostics. *Proc. IEEE Information Visualization Symp. (InfoVis)*, pp. 157–164, 2005.

[64] G. Wills and L. Wilkinson. Autovis: automatic visualization. *Information Visualization*, 9(1):47–69, 2010.

[65] T. D. Wilson, C. E. Houston, K. M. Etling, and N. Brekke. A new look at anchoring effects: basic anchoring and its antecedents. *Journal of Experimental Psychology: General*, 125(4):387, 1996.

[66] E. Yong. Nobel laureate challenges psychologists to clean up their act. *Nature*, 490:7418, 2012.