Computers & Graphics (2020)

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

On the perceptual Influence of Shape overlap on Data-Comparison using Scatterplots

Christian van Onzenoodt[a,*], Anke Huckauf[b], Timo Ropinski[a]

[a] *Visual Computing Group, Ulm University, Germany*
[b] *Institute of General Physiology, Ulm University, Germany*

ARTICLE INFO

Article history:
Received 1 August 2020

Keywords: Scatterplots, perceptual study, crowdsourcing

ABSTRACT

Scatterplots can be used for a wide range of visual analysis tasks, for example comparing correlations or variances of clusters across potentially multiple classes of data, in order to find answers to higher-level questions. Comparing classes of data in one scatterplot demands additional visual channels to encode this dimension. While perception research suggests colors as rather perceptually dominant, other studies show that shapes can also be visually salient. However, with an increasing amount of data, overlapping shapes can cause perceptual difficulties and obscure data. Even though shapes in scatterplots have been investigated extensively, the overlap between these shapes has usually been avoided by using synthetic scatterplots. To overcome this limitation, we investigate the perceptual implications of overlap when comparing data using scatterplots using a series of crowd-sourced user studies. These studies include common visual analysis tasks, like comparing the number of points, comparing mean values, and determine the set of points that is more clustered. To support our investigations, we introduced and compared four metrics for overlap in scatterplots. Our results provide insight into the overlap in scatterplots, recommend combinations of shapes that are less prone to overlap, and outline how our metrics could be used to optimize future scatterplot design.

1. Introduction

Scatterplots are widely used to visually explore and communicate abstract data. Oftentimes this involves dimensions of ordinal data, defining classes which observers would like to compare against each other. For this comparison, the data is often presented in a single scatterplot. But the ability to compare classes in a single scatterplot depends on the given data. As the data becomes more similar for example in terms of variance, or with an increasing number of points, we obtain a larger amount of overlap between points.

This overlap of data points can hide data or introduce artifacts in certain arrangements of data points. Therefore, overlapping data points influence the perception of an observer and can thereby hinder the ability to find answers to an analytic question. Additionally, the given task plays an important role, since tasks like finding outliers in a set of points might not suffer from high amounts of overlap, while it is difficult to identify clusters under such conditions. This leads to a need for optimization for a plot with respect to visual encodings, such as used shapes and their size, to enable observers to explore the data and find answers to analytic questions.

Although there is research on how to optimize scatterplot design in terms of overdraw, existing approaches are limited on for example only adjusting marker opacity [1]. Micallef et al. [2] introduce an approach to optimize marker size and marker opacity by using a cost function but limit their approach

*Corresponding author:
e-mail: christian.van-onzenoodt@uni-ulm.de (Christian van Onzenoodt)

to using colors to encode different classes. Other research intended to optimize scatterplot design with respect to the intended tasks [3, 4, 5], or by using animation to try to alleviate overdraw by constantly redrawing points over existing ones [6].

Perception research suggests that color hue is a visually dominant channel [7, 8]. However, recent work shows that shapes can also be a viable choice to encode categorical variables in scatterplots [9, 10, 11]. Since these shapes need to be relatively large to be distinguishable, shape overlaps become likely, especially for datasets containing a large number of data points. Overlapping shapes may also result in perceptual difficulties, such as occluded data or false positives. These kinds of perceptual difficulties can also lead to situations where certain arrangements of shapes lead to the formation of artificial new shapes (for example multiple plus symbols forming a square shape) which can not appear when only using one kind of shape in different colors. While other 2D scatterplot parameters, such as shape size [12] and shape color [13] have been investigated extensively, the overlap between shapes in a scatterplot has, to our knowledge, not been considered in depth yet. Furthermore, we see a lack of a concrete measurement of overlap appearing in a scatterplot, especially with a focus on human perception. So to support our research on shape overlap, we investigated three different measurements for overlap and compared their ability to support the prediction of human perception.

Since previous work could show that results of perceptual experiments can be used to predict human perception [14, 15, 16], we conducted a series of six user studies to evaluate which of our metrics could serve as a valuable predictor. To do so, we used Amazon's Mechanical Turk platform (MTurk) since it offers a large and diverse pool of participants [17, 18, 19]. Since all of these participants are not in a typical laboratory environment, such crowd-sourcing studies nicely reflect the variety of conditions under which users would inspect a visualization. We selected three common tasks based on previous work related to 2D scatterplot interpretation for our experiments. These tasks include judgments of the comparison of the number of shapes, the comparison of variance of sets, and the comparison of average value. Within our experiments, we have investigated a set of six different shapes, two different sizes of shapes, and a broad variety of overlap conditions reaching from almost no overlap to heavy overlap. Finally, we show how our findings can be used to optimize future, unseen scatterplots to improve the ability to solve a given task.

The remainder of this paper is structured as follows. First, we discuss work which is related to our investigations in Section 2, before we discuss our overlap measurements in Section 3. Section 4 outlines the methods used in our experiments, followed by three sections presenting the results. Afterwards, we evaluate our metrics in Section 8 and present direct implications on scatterplot design in Section 9. Finally, Section 10 concludes the paper and outlines possible future extensions of our investigations.

2. Related Work

The perception of shapes and the contrast between shapes have been investigated in previous work [20, 21]. Additionally, the usage of shapes in two-dimensional scatterplots has been investigated [10, 11], but so far overlap between shapes has been avoided through synthetic data generation.

Shapes in 2D scatterplots. When using shapes to encode additional dimensions within 2D scatterplots, it is required that these shapes are visually separable. Demirlap et al. [20] introduced distance matrices for perceptual judgments called *perceptual kernels*. To derive these matrices they conducted a set of crowd-sourced experiments on the MTurk platform. Comparison tasks between shapes, colors, sizes, and combinations thereof were investigated by using five different tasks. These tasks included pairwise comparison using two different Likert scales, triplet ranking with matching, as well as discrimination, and spatial arrangements. Based on the obtained results, they propose a metric that encodes the perceptual distance between for example different shapes, colors, and sizes of shapes.

To investigate which pairs of shapes offer good visual separability, Tremmel conducted two experiments [21]. He compared filled and non-filled shapes, as well as different types of circles with, for example, crosses or dots in the center, together with a plus shape and an asterisk. For his experiments, he created different synthetic scatterplots containing two shapes. He further ensured that the individual shapes are not overlapping by requiring a minimum distance between each of them. One of the shapes appeared more often, such that the task was to find the shape which appears more often. Tremmel's results indicate that a combination of filled and non-filled shapes provide a good visual separability. But he also found that shapes with different numbers of terminators (for example four terminators of a plus ($+$), or six terminators of an asterisk ($*$)) can be separated well.

Li et al. [10] investigated the effect of shape and size on the perception of scatterplots. To do so they conducted a user study including different tasks, where they for instance estimated which shape appears more often, identified outliers, and determined which shapes are more clustered. Thus, for generating their test stimuli, they divided the scatterplot area into a uniform grid and placed the shapes into randomly selected cells to prevent overlapping shapes. Their results indicate a good visual separability between polygon- and asterisk-like shapes.

Finally, Burlinson et al. [11] could show, that shapes such as squares, triangles, circles, asterisks, plus- and cross-shapes can indeed be used for tasks like counting the number of elements and guessing the average of a cluster. They divided these six shapes into the classes *open* and *closed*, where squares, triangles, and circles are considered closed shapes while all the remaining are considered open shapes. To generate their plots, Burlinson et al. used an approach introduced by Gleicher et al. [22] which prevents overlapping shapes. Under these overlap avoiding conditions, they found that using two shapes from the same category causes a significant effect on response times as well as error rates when compared to using shapes from different categories. Furthermore, they found, that open shapes seem to be a better choice to be used as target shapes.

Overlap in 2D scatterplots. Mayorga and Gleicher [23] proposed *Splatterplots*, an enhanced presentation of scatterplots which tries to solve the problem of overlap. These *Splatterplots*

abstract the presentation of dense regions in scatterplots into color coded contours while keeping outliers as single points. *Splatterplots* work well even for millions of data points, but require interactions such as continuous zoom to reveal the actual data. Another technique of preventing overlap in scatterplots was developed by Keim et al. [24]. Their approach tries to minimize the overlap in a plot by distorting the scales of a two-dimensional scatterplot. The user can freely distort the plot until an optimal trade-off between distortion and overlap is achieved. While they propose techniques to automatically optimize plots, distorting the data can introduce unwanted relations.

Urribarri and Castro [25] recently developed a metric to measure the number of shapes that are not completely overlapped by other shapes. They show how to use this metric for different shapes and how a scatterplot can be optimized using their metric. However, the proposed metric only calculates the total amount of completely hidden shapes in a scatterplot. This is especially problematic when using filled shapes which are relatively large. Under this circumstance, the amount of completely hidden shapes becomes a problem. Furthermore, shapes that are only overlapping partially are not considered by their metrics, and the actual influence on the perception of overlap is also neglected.

Related works by Matejka et al. [1] and Micallef et al. [2] try to improve the appearance of scatterplots by adjusting the size and opacity of data points. While this approach does maintain the general distribution of points and does support users to find regions where points are really clustered, it is difficult to apply this approach when using different shapes to encode classes. As the overdraw increases, and therefore the opacity of individual points decreases, it becomes more and more difficult to distinguish between points.

Furthermore, Chen et al. [6] introduced an approach to overcome overplotting by using animation and iteratively redrawing points on top of existing ones. This makes cluttered regions in the plot easy to detect, but this approach does not work in environments where animations can not be used, for example for printed plots.

3. Overlap Calculation

To be able to estimate the degree of overlap present in a 2D scatterplot, an appropriate metric is needed in order to investigate the influence on human perception. These metrics should further serve as an additional predictive variable in our investigations and help to understand the general amount of overlap. Although we suppose that a more precise measurement leads to better predictive performance, we tried to find some viable simplifications which may also lead to viable predictive performance using our model. Therefore, we compared four different metrics using the acquired data from our user studies.

We have not considered the metric proposed by Urribarri and Castro [25], which measures the amount of data that is totally hidden by other shapes since it assumes that scatterplots use filled shapes. Furthermore, the goal was to focus on overlap instead of hidden data, which can nicely be done by the four metrics discussed in the following paragraphs. Also, algorithms which depend on convex hulls are not viable, since we then would need to come up with an artificial hull for our open shapes. Therefore, we also decided against different kinds of collision detection algorithms, which rely on convex hulls, like the algorithm introduced by Gilbert et al. [26].

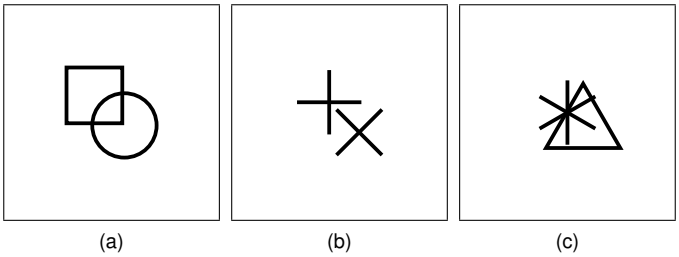

Fig. 1. Examples of overlap appearing between different combinations of two shapes. Shapes are drawn with a pixel size of fifteen and a line width of one. The overlap is measured by using all three of our introduced overlap measurements, where M_{pix} measures on a pixel based level, M_{num} counts the number of overlapping shapes based on a bounding circle, and M_{rel} additionally takes into account how much these bounding circles overlap. This would produce the following overlap measurements: (a) M_{num} = 1.0, $M_{rel} \approx$ .29, $M_{pix} \approx$ .02, $M_{shape} \approx$ .02; (b) M_{num} = 1.0, $M_{rel} \approx$ .29, M_{pix} = 0, M_{shape} = 0; (c) M_{num} = 1.0, $M_{rel} \approx$ .67, $M_{pix} \approx$ .12, $M_{shape} \approx$ .55.

3.1. Discusssion of Metrics

Our goal was to find a measurement for overlap in scatterplots. Therefore, we decided to use a pixel-precise measurement as our baseline metric, before trying to find other metrics which might reflect human perception in a better way. The most precise way of measuring overlap in scatterplot would be to just count the number of pixels of a shape that is overdrawn by other shapes.

While we hypothesize that this is the most precise way of measuring overlap, we also suspect that this metric does not serve as a good measurement of overlap on a perceptual level. For example, when two open shapes like the asterisk ($*$) and the plus ($+$) overlap each other, there are cases where the shapes are not overlapping on a pixel level, but humans might perceive these shapes as overlapping and having trouble in separating the shapes. Therefore, we came up with a second metric that measures the overlap between shapes using bounding circles, where we just count the number of overlaps between the bounding circles. We choose the size of the bounding circles to be of the size of the shape. This means for a square ($\square$) with a width and height of seven pixels, we used a bounding circle with a radius of seven pixels. While this introduces false positives, we suspect this metric to reflect the human perception in a better way than the pixel precise measurement, especially when shape size decreases. Figure 1b shows an example of such a false positive measurement.

This second metric, however, does not take into account how much two shapes are actually overlapping. To overcome this limitation, we came up with a third metric which does include these criteria. In doing so, a plot where all of the shapes are just slightly overlapping would produce a different amount of measured overlap, when compared to a plot where all the shapes are overlapping a lot.

So in the end, we came up with the following metrics:

M_{pix} **: Pixel-based overlap.** For this metric we rendered the scatterplot and the shapes appearing in our plots and counted the number of pixels for the individual shapes. We then used the plots generated for the conducted experiment, where we know how many points are shown and which shape they are using. Using this information we can compute how many pixels in a given plot should be occupied by shapes. Afterwards, the number of used pixels in the scatterplot can be used to calculate the number of pixels that appear to be overdrawn. Since our closed shapes are drawn to be transparent in the center, such a shape only occupies the number of pixels used for the outline of the shape.

M_{num} **: Number of bounding circle overlaps.** For this metric, we simply count the number of overlapping shapes by using a bounding circle for each shape. To do so, we calculate the distance between each of the data points to all other points while excluding already compared pairs. If the distance between two points was smaller than the size of a shape, we found an overlap. In the end, we normalized this number by dividing through the total number of possible overlaps.

M_{rel} **: Relative bounding circle overlaps.** We created a special version of our bounding circle overlap test (M_{num}), in which we also calculated to which degree the bounding circles overlap. Calculating the percentage of overlap between each individual shape offers the possibility to have a more precise measurement of interactions between shapes. We again compare each point with each other (excluding already tested pairs). If we find an overlap, we compute the percentage of overlap by dividing the distance through the size of the shape and subtract this from one. To compute the final result for the complete plot, we calculated the mean of all the overlaps measured.

M_{shape} **: Shape-based overlap.** For our final metric, we intended to focus on the overlap between individual shapes and the way they interact with each other. We, therefore, rendered all combinations of all shapes and used one shape as a sliding occluder for the other shape in a way that we ended up with all possible (pixel-precise) constellation between the two. Afterward, we calculated the number of overdrawn pixels for each of these constellations so that we end up with a value of overlap depending on the relative position between the two shapes. These values are then normalized from the minimum (which is always zero) and maximum possible overlap for a pair over all constellation to the range of zero and one. Finally, to use this metric for our scatterplots, we computed the overlap for the complete plot and calculated the mean.

Figure 1 shows a comparison of overlap when measured with the four proposed metrics, and Figure 2 presents overlap measurements from four of our used stimuli.

4. General Methods

To investigate the perceptual influence of overlap in scatterplots, we conducted six experiments. To be broad wrt. shape and task, we have used three different visual analysis tasks and two different shape sizes. The general approach which all of the experiments have in common is described in the following sections.

4.1. Task Selection

For our investigations on the effect of overlap, we selected tasks with a focus on the comparison between classes from the task definitions for scatterplots by Sarikaya and Gleicher [3], which have also established by previous work. From their classification, we decided to use tasks from the *aggregate-level* category. This category describes tasks, which are common when answering higher level questions by aggregating sets of data points. We did not use tasks that are based on finding outliers since they are not prone to overlap. Furthermore, we decided against using a task involving the comparison of correlation.

We found, the task of finding and comparing average values is easier to communicate in a crowd-sourced environment, where people do not necessarily have an understanding of the abstract concept of a correlation. Furthermore, we argue that perceiving correlations is less prone to overlap since it just requires the perception of the outer shape of a set of points, rather than perceiving individual shapes.

Besides that, our goal was to use tasks which do enable participants to answer a given question rather quickly, rather than having to investigate the plot over longer periods. Thus, we decided to use the following tasks.

Comparing Number of Shapes. Within this task, users are confronted with 2D scatterplots containing two types of shapes, and they have to decide which shape appears more often. This task does not only fulfill all our criteria for a large scale user study, but it has also been extensively used in previous work [12, 11, 10, 21], which makes our results transferable.

Comparing Variance. During the second task, users are also confronted with 2D scatterplots containing two different shapes, but now they have to determine which set of shapes is more clustered and thus has a smaller variance. Again, this task fulfills all our task requirements and has been used in previous research [12, 10]. Like in previous work, we also choose the more clustered shape as target, since we suspect this shape to suffer more from overlap than the shape with the larger distribution.

Comparing Average Value. Within the third task, we ask the study participants to judge which of the two displayed sets of shapes has a higher average y-coordinate. Again, the task involves the comparison between two sets of shapes and has also been used in previous work [22, 11]. Furthermore, we decided to use this task, since, in contrast to the first two tasks, this task involves *visual aggregation* in a way that the observer computes the aggregated properties over a collection of points. Such an aggregation and the comparison between the results of such aggregations are common as it corresponds to many decisions like if there is one class in the data that is better than another. However, in contrast to the first two tasks, this task is rather complex. We, therefore, included additional control questions during our user study to ensure the quality of the results.

4.2. General Stimuli Generation

Previous work suggests that a combination of open and closed shapes works well in terms of distinctiveness [11]. Based on this finding and the distinctiveness of shapes predicted by Demiralp's perceptual kernels [20], we decided to include the following six shapes: circles (○), squares (□), triangles (△),

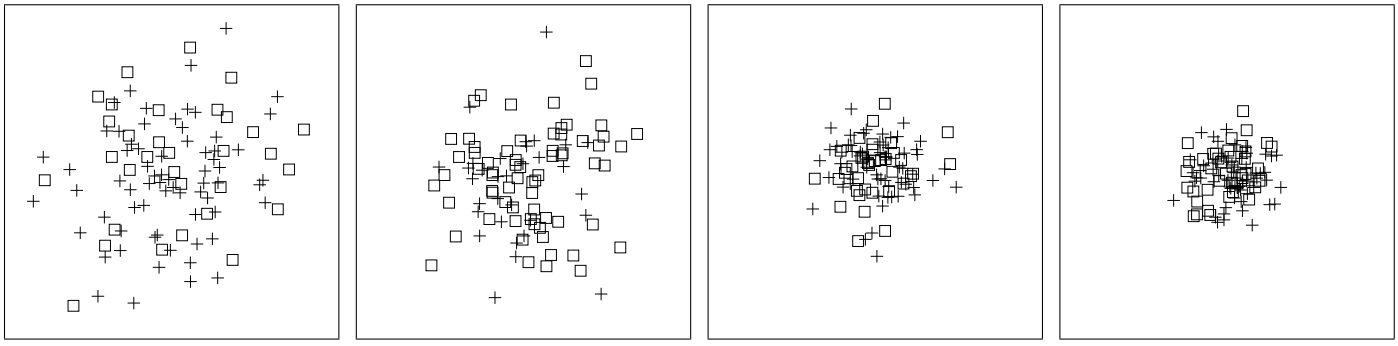

Fig. 2. Comparison of different amounts of overlap, appearing in our scatterplots. (a) shows an example of one of the lowest overlap measurements ($M_{num} \approx .0097$, $M_{rel} \approx .0025$, $M_{pix} \approx .0089$) we used in Experiment 1, while (d) shows an example of one of the highest overlap measurements ($M_{num} \approx .062$, $M_{rel} \approx .0211$, $M_{pix} \approx .1548$). (b) shows medium low overlap with $M_{num} \approx .017$, $M_{rel} \approx .0055$, $M_{pix} \approx .0561$, while (c) shows medium high overlap with the following measurements: $M_{num} \approx .049$, $M_{rel} \approx .0162$, $M_{pix} \approx .1202$.

crosses (×), pluses (+), and asterisks (∗). While circles (○), squares (□) and triangles (△) are considered to be closed shapes, crosses (×), pluses (+), and asterisks (∗) are open shapes. Each shape was presented as a target shape together with one distractor shape, whereby we have realized all possible combinations thereof.

Even though we think that using more than two shapes together in one plot would lead to interesting effects, for example when using cross (×), plus (+), and asterisk (∗) together, we decided against going beyond a two-way comparison by adding another distractor class. Using more than two shapes at a time (and all the combinations thereof) would open up a space of combinations that would go beyond what would be manageable in a user experiment. This is especially true considering all the other parameters we would like to investigate during our experiments.

During the study, each of the scatterplots showed a total of 100 data points divided into two sets where each set was encoded using a different shape. We chose normal distributions to generate our pointsets because these are frequently used as they underlie many natural phenomena. Although we used a normal distribution for all our pointsets, the average task, as well as some combinations in the other tasks, use a distribution with a wider spread where the resulting pointset spreads evenly over the canvas. In doing so we cover a wide range of different point distributions ranging from even spreading over the canvas down to heavy amounts of overdraw. The normal distribution was generated by using a pre-defined seed, to ensure the reproducibility of our pointsets. We further verified that in all generated pointsets all of the shapes are completely shown on the canvas, such that they are not clipped by the border.

We further chose to use variance as a helper to generate different amounts of overlap indirectly. Rensink and Baldridge used a method of generating just-noticeable-difference staircase approach to generate stimuli to investigate correlations in scatterplots [27]. While this is a viable approach which could have been adopted to generate different amount of overlap, we decided to use fixed amounts of variances to generate our stimuli. The main reason for this is, that we already include a rather large number of variables in our experiments, and we also repeated each target-distractor condition using each variance with a different seed. By doing so, we also generated a continuous variance in overlap appearing in our stimuli. Besides that, our focus was to find combinations which are less prone to overlap. Having combinations of shapes that show promising results in our studies, could then be used in further studies, using such a stair-case approach to further investigate the interaction of overlap and shapes.

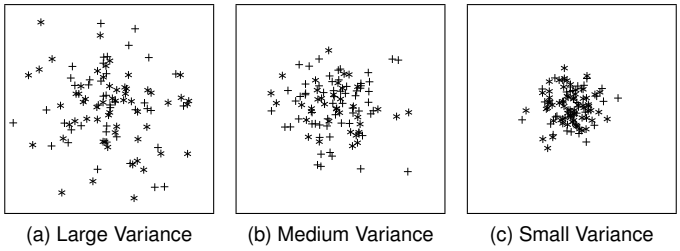

Fig. 3. Comparison of variances used in Experiment 1 (count task, big shapes). (a) shows an example of the large variance for both shapes. (b) shows an example of the medium variance for both shapes, while (c) shows the small variance for both shapes.

In general, we used three different variances, that were chosen in a way that the largest variance produces plots in which the shapes are almost evenly distributed over the entire canvas. The smallest variance produces distributions that are just big enough such that the shapes are still visible. The third variance was chosen to be in between the smallest and the largest one. Concrete values of means, covariances, and seeds for the random generation, used throughout our experiments can be found in the supplementary material. Figure 3 shows a comparison of our used variances.

The closed shapes (○, □, and △) were drawn with a transparent center because filled shapes would introduce an additional variable to reflect the order in which the different shapes are drawn, which directly affects the overlap. Furthermore, drawing open shapes, like for example a plus, in front of a filled closed shape, for example, a circle would end up in the same re-

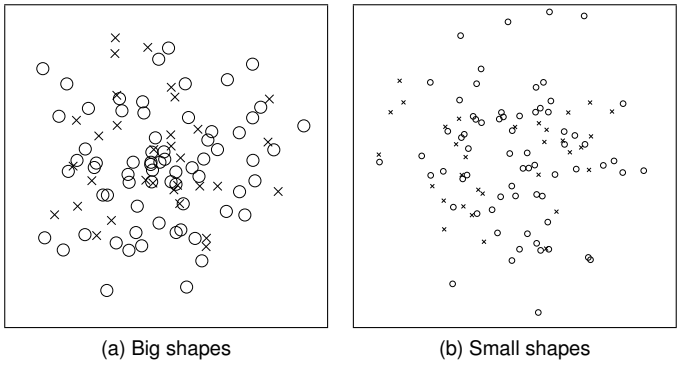

Fig. 4. Comparison of shape sizes used in our experiments. (a) shows an example of our big shapes with a pixel size of fifteen pixels, while (b) shows an example of the small shape size of seven pixels. Both examples are taken from our experiment involving the comparison number task. For both examples circle is the target shape while cross is the distracting shape.

sult as using transparent closed shapes. All shapes were drawn with the center of mass on the actual data point. This means a triangle is slightly shifted into the positive y-axis when compared directly to for example a square or a circle. Also, the shapes are drawn so that they are about equal in terms of area. This means when comparing a square to a circle, the circle takes up slightly more space along both x- and y-axis. The open shapes (+, ×, ∗) were drawn in a way so that they would fill the circle shape with their line endings and therefore also take up an equal amount of space. This was done to ensure that each of the shapes had a similar strong perceptual impact and this way was about equally salient. The outlines of our shapes were drawn with a line width of one, while the canvas was chosen to be white and 400 by 400 pixels in size, as it should work on all modern desktop computers and was also used in previous studies [22].

For each of the tasks described above, we conducted two experiments. In a first experiment, we used a pixel size of fifteen pixels for the shapes and in a second experiment, we used a pixel size of seven pixels. These sizes were chosen, since seven pixels mark a lower limit in terms of usability, while fifteen pixels mark an upper limit. If we would draw shapes smaller than seven pixels in size, the square (□) and (○) become hard to distinguish. Also, at a shape size smaller than seven pixels, the asterisk (∗) is starting to become a filled square symbol and does almost take up the complete area. Shapes bigger than fifteen pixels also do not seem usable when drawn on a canvas with 400 pixels in size. Also, fifteen pixels have been chosen, since it doubles the size of the small shape size while still having an exact center for the shapes since fifteen is an odd number. Figure 4 shows a comparison of our used shape sizes.

To draw the plots for our online survey, we used *Data-Driven Documents* (D3) [28], which uses the ability of the browser to render SVG images, and thus can be used to generate vector-based plots. The Figures 2, 3, 5, 6, and 7 show examples of the used plots.

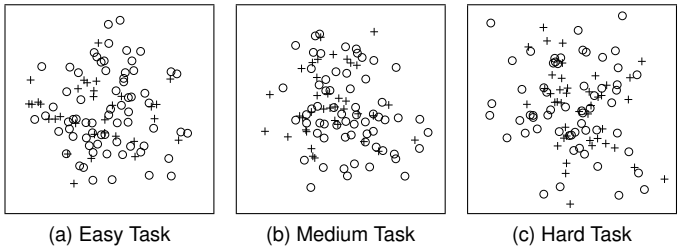

Fig. 5. Examples of stimuli, as used in Experiment 1 (count task, big shapes). For all shown plots, circles are the target shape, while pluses are the distractor. The easy task shows 68 circles and 32 pluses, the moderate task shows 63 circles and 37 pluses, and the hard task shows 58 circles and 42 pluses.

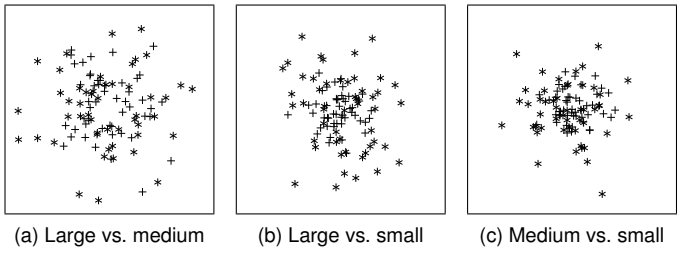

Fig. 6. Examples of stimuli, used in Experiment 3 (variance task, big shapes). For all examples, asterisk is the target, while plus was used as a distractor. (a) shows an example where the target shape uses the large variance and the distractor shape uses a medium variance as described in Section 4.2. (b) shows a large variance for the target shape and a small variance for the distractor, while (c) uses our medium variance for target shapes and small variance for distractor.

4.3. Task Based Experimental Design

Based on the general method for stimuli generation, we have generated scatterplots for all three tasks. Within this subsection, we describe how these stimuli vary wrt. task, and discuss the combinations of parameters as used in our user studies.

Comparing Number of Shapes. As described in Section 4.1, in this task participants had to rate which shape appears more often in a scatterplot. We divided 100 data points into two groups with different deltas between the groups to generate easy, medium and difficult tasks. The actual deltas have been adopted from Burlinson et al. [11]. So for easy tasks, we used 68 shapes for the target set and 32 distracting shapes, for the moderate task we used 63 target shapes and 37 distracting shapes, and for the hard task, we used 58 target shapes and 42 distracting shapes. To generate different amounts of overlap between shapes, we used three different variances as described earlier. We used the same variances for both, the set of target and the set of distracting shape and also for both the x- and the y-axis. For all of the normal distributions, the mean was chosen to be in the center of the canvas. Using each of the shapes as a target (6), all other shapes as distractor (5), three different task difficulties (3), and three different variances (3), we created 270 different combinations. For each of these combinations, we created three scatterplots, where each of them uses another seed for the normal distribution. This way we were able to have the same combinations with different amounts of overlap and added repetition to our stimuli combinations. By doing so, we created

a total amount of 810 stimuli for the comparison of number of shapes task. Figure 5 shows examples of stimuli used for this task.

Comparing Variance. Within this task, participants must judge which set of shapes shows a smaller variance and therefore is more clustered. We again divided 100 data points into two groups but used two equally sized groups for this task. To generate different task difficulties we used three different combinations of variances. Large vs. medium variance to generate a plot with a low amount of overlap, medium vs. small variance to generate a plot with a high amount of overlap, and large vs. small variance to generate a medium amount of overlap. The mean of the normal distribution was chosen to be in the center of the canvas. Using each of the shapes as target (6), all other shapes as distractor (5), and our variance combinations (3) we generated 90 combinations. We again repeated each combination using three different seeds for the normal distribution to generate a total of 270 stimuli for the comparing variance task. Figure 6 shows examples of stimuli used for this task.

Comparing Average Value. In this task, participants were asked to judge which set of shapes was on average higher in the y-axis. Again we used 100 data points, divided into two equally sized groups. Burlinson et al. used a dart-throwing approach [29] to generate datasets without overlapping shapes which they used as stimuli in their user study. Since our experiments focus on overlap, we instead also used a normal distribution to generate stimuli for this task. To achieve a uniform distribution of the points over the complete canvas, we used a large distribution for both of the sets. Different task difficulties are generated by adopting the approach of Gleicher et al. [22] where the distance along the y-axis between the means of two sets is measured. This distance parameter is called Δ. We also adopted the Δ values reported by Gleicher et al.: 8, 16, 24, 32, 40, and 80 as used for control questions. We used different means for the normal distribution to generate pointsets with the desired amount of offset along the y-axis between the clusters. After generating the pointsets, we ensured that we obtain the correct distance between the sets by calculating the given average and offsetting the points to the desired distance in averages. So by using all of the shapes as target (6), all others as distractor (5), and our Δ values (6) we have generated 180 combinations, resulting in 540 stimuli for this task, as we again use three different seeds for each combination. Figure 7 shows examples of stimuli used for this task and compares the different Δ values used to vary difficulty.

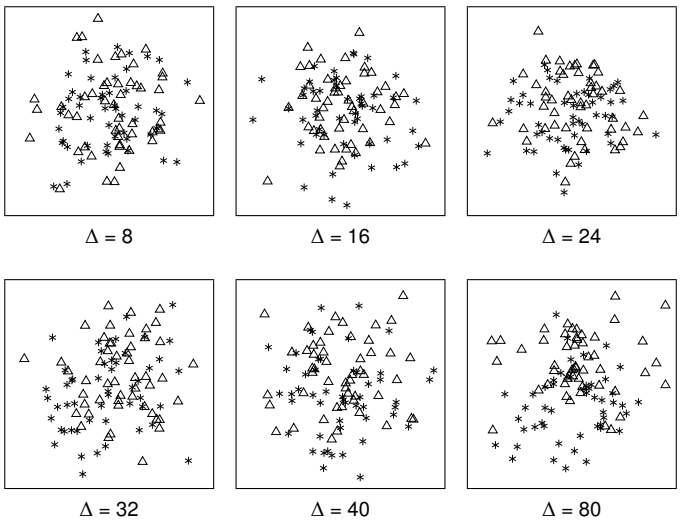

Fig. 7. Examples of stimuli, used in Experiment 5 (average task, big shapes). For all plots, triangles are the target shape, while the asterisk was used as a distractor. Task difficulty was varied using different distances between averages (Δ), reaching from 8 pixel to 80 pixel. For all examples, the triangle was used a target shape, while the asterisk symbol was used as a distractor.

4.4. Procedure

Each of the conducted experiments, started with a demographic questionnaire, followed by an introduction to the task. This introduction included five examples of plots similar to the ones used in the study with an explanation of the task. After the introduction, the participants completed five practice trials to get used to the tasks. The stimuli for the introduction as well as the practice trials were generated by using configurations which were also used in the study, but using custom seeds to create the normal distribution. This way we ensured that none of the example or training stimuli appear in the study.

For each of the stimuli, the participants had to answer by pressing the "f" or "j" key, such that they could rest their hands on the keyboard comfortably. Each of the keys showed the shapes assigned to the key, which were randomly assigned to one of these keys. The participants were instructed to respond to the stimuli as quickly as possible while still making sure to give the correct answer. By instructing to answer as fast as possible, while still making sure to give the correct answer, we tried to make participants answer based on their intuitive decisions.

In our introduction, we explained our task before giving examples including an explanation of the shown plot and a hint towards the correct answer. If a participant answered incorrectly, we showed additional explanations instructing the user to press the correct key. Before the stimuli were presented, a fixation screen containing a plus shape in the center of a white canvas was shown for a random time of 500, 600, 700, 800, 900, or 1000 milliseconds. We adopted this approach from Burlinson et al. [11] to prevent the participants getting used to the timing and just clicking through the survey. The response time was restricted to ten seconds to prevent the user from solving the task by for example counting the shapes on the screen. If a participant exceeded this time restriction, the response was discarded and a red error screen was shown instructing the user to answer more quickly.

4.5. Participants

Over all our studies we recruited a total of 624 participants (258 female, 360 male, 2 other, 4 did not report, $M_{age} = 33.70$, $SD = 10.55$). We had to exclude a total of 46 participants due to poor performance in terms of accuracy, or failing in control questions as used in the study. To compensate for learning effects, for each of the six individual experiments we excluded all participants from previous experiments.

5. Comparing Number of Shapes Experiment

To investigate the perception of the comparing number of shapes, we have conducted two experiments with two different shape sizes. Within this section, we first describe the methods used in these experiments, before analyzing the results.

5.1. Methods

As described in Section 4.2, we used big and small shapes in our studies, which is also the difference between the two experiments conducted for this task. Thus, in Experiment 1 we used big shapes, while Experiment 2 used small shapes. The procedure of both experiments follows the description provided in Section 4.4. To reduce the workload of the individual participants, we divided all 810 stimuli randomly into ten groups, so that each participant had to rate 81 stimuli. This way we limited the length of each survey to about 10 to 15 minutes to ensure quality and motivation of participants [18].

In Experiment 1, where we used big shapes with a pixel size of fifteen pixels, 115 participants took part. We had to discard the responses of four participants, due to high error rates (> 50%). Thus, we analyze data from 111 participants (52 female, 58 male, 1 did not respond).

For Experiment 2 we used small shapes with a pixel size of seven pixels. We excluded all participants of Experiment 1 from this experiment. In Experiment 2, we acquired data from 110 participants and had to exclude five participants because of high error rates (> 50%). Therefore we present the results of 105 participants (42 female, 63 male).

5.2. Analysis

After providing an overview of the acquired user feedback, we analyze which combinations of shapes resulted in the best accuracy within this section.

5.2.1. Data Description

Overall participants showed a mean accuracy of 73.57% (SD = 8.80%) of correctness answers for Experiment 1 and 71.89% (SD = 9.54%) for Experiment 2. Only 0.25% of the trials ran into the time out for Experiment 1 and 0.29% for Experiment 2. Response times over participants were lower in the experiment using the small shapes (M = 1770.88ms, SD = 1293.97ms) when compared to the experiment using the big shapes (M = 1915.04ms, SD = 1300.42ms). We hypothesized this effect, because larger shapes result in a larger amount of overlap, making the task more difficult. Using Friedman's ANOVA we tested the effect of intended task difficulty on participants' accuracy and found a significant effect for both experiments (Experiment 1: $\chi^2(2) = 67.12$, $p < .001$; Experiment 2: $\chi^2(2) = 71.93$, $p < .001$).

5.2.2. Shape Effects

To analyze the effects of individual shapes, we compared all combinations of shapes used in both experiments using this task. Figure 8 shows a comparison between a selection of the best and the worst combinations in terms of accuracy. Without any further statistical analysis, we found that there are combinations that seem to work much better than others. Our measured accuracies for combinations of target and distractor shape varied between around 90% accuracy for our best combinations (○ / +, △ / +) down to around 50% for some of the worst combinations (□ / △, + / △). While these results overall indicate that findings of Burlinson et al. [11] also hold for conditions where shapes overlap, there are also interesting outliers like the triangle (△) vs. square (□) or asterisk (∗) vs. plus (+) combinations which exhibited very high accuracies under the comparing number task.

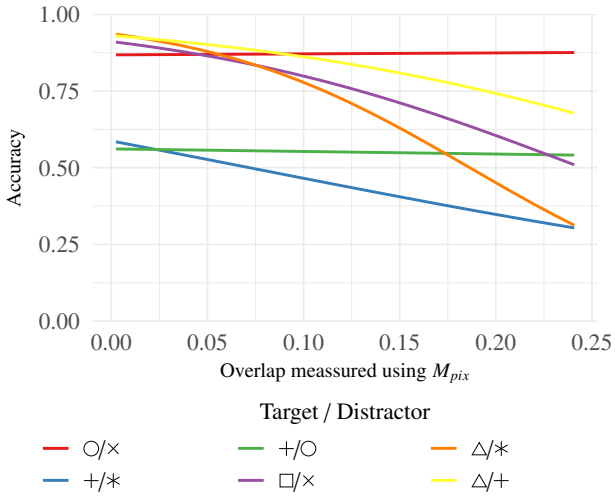

Fig. 8. Accuracy while overlap increases for the experiments involving the comparing number of shapes task. We show logistic regression curves for a selection of shape combinations which showed overall best and worst performance. Overlap was measured using our pixel-based metric M_{pix} .

Furthermore, we investigated which combinations of shapes suffered stronger from overlap than others. In Figure 8, we present a selection of combinations for target and distractor shapes as used in this task. The figure shows logistic regression curves of accuracy for these combinations.

The combinations have been selected based on if they either showed good or bad accuracies in this task, as also shown in Table 1. Additionally, we selected combinations where the increase of overlap showed a strong incfluence on accuracy. The regression lines in Figure 8 shows that for example the combination of triangle (△) and asterisk (∗), as well as plus (+) and asterisk (∗), suffer drastically from an increasing amount of overlap. In these cases, the accuracy drops below the accuracy of chance, indicating that the asterisk shape has a strong influence when used as a distractor. This influence seems to indicate that the asterisk makes the set appear to have more points than it has.

We investigated if our used shape sizes had an effect on participants' performance in terms of accuracy using Wilcoxon rank-sum test. This test was used since our accuracy data failed a statistical test on normal distribution. When comparing the results of our experiment using the number of shapes, we found no significant difference between our experiments (W = 6454.5, p = .17).

6. Variance Task Experiment

Using this second task we investigated the influence of overlap on the estimation of variance. Again, in a first experiment (Experiment 3), we used big shapes while in the second experiment (Experiment 4) we used small shapes as described in Section 4.2. We again first describe our used methods, before discussing our results.

6.1. Methods

The experiments for the variance task was also set up as described in Section 4.4. The generated 270 stimuli were randomly divided into five groups, so each participant had to rate 54 stimuli. We again conducted two disjunct experiments, where all participants of previous experiments where excluded from subsequent experiments. Experiment 3 (using big shapes) has been conducted by 58 participants, of which we had to exclude eight participants because of high error rates (> 50%). The experiment using the small shapes (Experiment 4) has been conducted by 68 participants. Here we had to exclude five participants because of high error rates (> 50%). So we present data of 50 participants (24 female, 24 male, two preferred not to answer) for Experiment 3, and 63 participants (28 female, 35 male) for Experiment 4.

6.2. Analysis

We again describe the overall performance of our participants before analyzing which shapes appear to work better for the given task.

6.2.1. Data Description

For this task, our participants overall showed the best accuracies when compared to the other conducted experiments. Over our two experiments, participants show almost exactly the same performance for the experiment using the small shapes (M = 79.19%, SD = 11.12%) and for the big shapes (M = 79.18%, SD = 13.24%). Among all responses, .34% of the trials timed out using the big shapes, and .27% using the small shapes, while the response times were lower using the big shapes (M = 1919.26ms, SD = 1393.95ms) as compared to when using the small shapes (M = 2163.56ms, SD = 1466.10ms). We used three different combinations of variances to generate tasks with different difficulties and amount of overlap. We found that these combinations of variances as shown in Figure 6 had a significant effect on accuracy for both of our experiments using this task (Experiment 3: $\chi^2(2) = 39.46$, $p < .001$; Experiment 4: $\chi^2(2) = 41.13$, $p < .001$).

6.2.2. Shape Effects

Investigating our results wrt. used shapes, we again found large shape dependent accuracy differences. As for our experiments using the comparing number task, we found that the combination of asterisk (∗) and square (□) showed the best accuracy (92.6%). Using the same combination of shapes, but square as target and asterisk as distractor, however, showed one of the worst performances in terms of accuracy using the variance task (71.2%). Also, we found again that the asterisk (∗) vs. plus (+) combination works really well in terms of accuracy with 88.6% of correct answers using this combination, indicating that there are indeed some combinations which seem to be an outlier from the open and closed categories [11]. Figure 9 shows a comparison of a selection of combinations of target and distractor shapes using this task. As in Section 5, we again, present combinations that showed good or bad accuracies for this task. Investigating combinations of shapes regarding overlap, we found that in this experiment overall overlap was a less stronger factor when compared to the count task based experiments. We suppose these results to happen because stimuli in this task overall showed less overlap since one of the sets needed to show at least a medium large variance. We again compared both of our experiments using the Wilcoxon ranked-sum test, and again found no significant difference on participants' accuracy when using different sizes of shapes (W = 1613.50, p = .83).

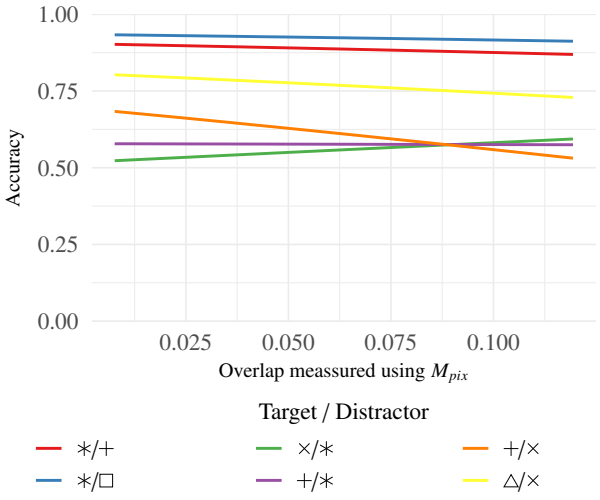

Fig. 9. Accuracy while overlap increases for the experiments involving the variance task. We show logistic regression curves for a selection of shape combinations which showed overall best and worst performance. Overlap was meassured using metric M_{pix} .

7. Average Task Experiment

For this final task, we also conducted two experiments using two different shape sizes. In the following subsections, we again first present an overview of the used methods before discussing our results.

7.1. Methods

The experiments for the average task were also set up as described in Section 4.4. Again, we conducted two experiments, one using big shapes (Experiment 5) and one using small shapes (Experiment 6). We randomly divided our 540 stimuli into nine groups, so that each participant had to rate 60 stimuli.

Again we excluded all participants from our previous experiments, as well as participants of Experiment 5 for Experiment 6. The experiment involving the big shapes (Experiment 5), was conducted by 145 participants, while 137 participants conducted Experiment 6 investigating the small shapes. For this task, we had to sort out a relatively large number of participants

as they failed in our control stimuli. Out of the 145 participants in Experiment 5, 34 participants failed in more than 50% of the control stimuli, while for Experiment 6, 47 out of 137 participants failed this condition. However, we suspect participants failing the control questions did not understand the task well enough to solve it. We suspect this happens since these tasks are harder to understand and requires a deeper understanding of the plot when compared to the first two tasks.

Even though we needed to exclude a large number of participants, we could still ensure that at least 10 participants were acquired for each of our nine groups. Therefore, we present the results of 111 participants (40 female, 70 male, 1 did not respond) for Experiment 5, and 90 participants (39 female, 51 male) for Experiment 6. Again, for further analysis, the responses to our control stimuli were excluded.

7.2. Analysis

We first present an overview of our acquired data before comparing which combinations of shapes showed the best performance for this task.

7.2.1. Data Description

Participants showed the worst performance in terms of accuracy using this task. We suppose this to happen since this task involves a higher level of understanding of the data when compared to the judgment of how many shapes are shown, or which shape has a wider spread over the canvas. For Experiment 5, participants showed a mean accuracy of 69.07% (SD = 13.41%), while for Experiment 6 the accuracy was with 66.78% (SD = 15.12%) even lower. However, time outs on this task were low again (0.37% for big shapes; 0.29% for small shapes), indicating that the time restriction was appropriate. For this task participants also showed the highest response times when compared to the other tasks, which again indicates that this task was more difficult. In our experiment using the big shapes the response times where higher (M = 1991.32ms, SD = 1580.62ms), when compared to using the small shapes (M = 1783.34, SD = 1312.37ms). We again tested, if the used difficulties had an influence on participants accuracy and found significant effects for both of our experiments (Experiment: 5 $\chi^2(4) = 112.93, p < .001$; Experiment 6: $\chi^2(4) = 39.213, p < .001$).

7.2.2. Shape Effects

We again compared how the used shapes affected accuracy. The combination of circle (○) and plus (+) again showed the best accuracy slightly outperforming the combinations triangle (△) vs. asterisk (∗) and square (□) vs. plus (+). The choice of the target again showed a strong effect, since the combination of plus (+) as target and circle (○) as distractor was one of the worst combinations with only 56.7% of correct answers. Compared to the two previous tasks, we found that the open and closed categories seem to work really well for this task since the seven best combinations are a combination of a closed shape as target and an open shapes as distractor. As we used a large variance for both of the pointsets (target and distractor), this task involved less overlap when compared to the count task. However, we found that for this task essential the combination of triangle (△) and cross (×) suffered from an increasing amount of overlap. Figure 10 shows a comparison of some of the best and worst combinations of shapes in terms of accuracy for this task.

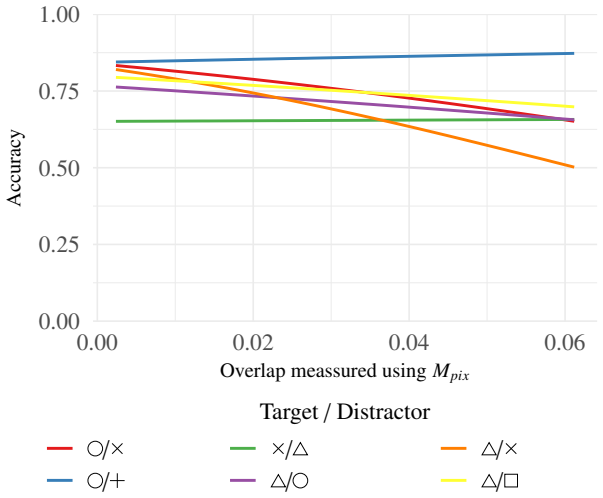

Fig. 10. Accuracy while overlap increases for the experiments involving the average task. We present logistic regression curves for a selected combinations of target and distractor shapes which overall showed good and bad performance in terms of accuracy. Overlap was meassured using metric M_{pix} .

As with our previous experiments we again compared participants' accuracy between both of the used shape sizes and found no significant difference (W = 5389.50, p = .3366).

8. Regression Model

To evaluate the proposed overlap metrics, and to investigate how they can be used as a predictive variable on unseen scatterplots, we used a regression model to fit our data. Since the outcome of correctness is binary and therefore limited to two discrete values, we used a logistic regression model. Such a logistic regression can be used to model the probability of a binary event (e.g. observers' ability to solve a given task) based on a set of given parameters (e.g. visual parameters of a stimulus). The predictive performance of different models (based on which parameters are used in the model) can then be compared to find which set of parameters describes the data the best.

To compare the models we first defined a null model without any predictive variables and subsequently added more predictive variables (target shape, distracting shape, and other dependent variables such as for instance amount of shapes). For each of these predictive variables, we determined if they can improve the predictive performance of our model by using the likelihood ratio test. After we defined this model we then added each of our metrics to this model and verified if they can further improve the predictive performance, again by using the likelihood ratio test. If the metrics improved the model, we then compared which of the metrics improved the model the most. This is done by comparing two models containing different metrics

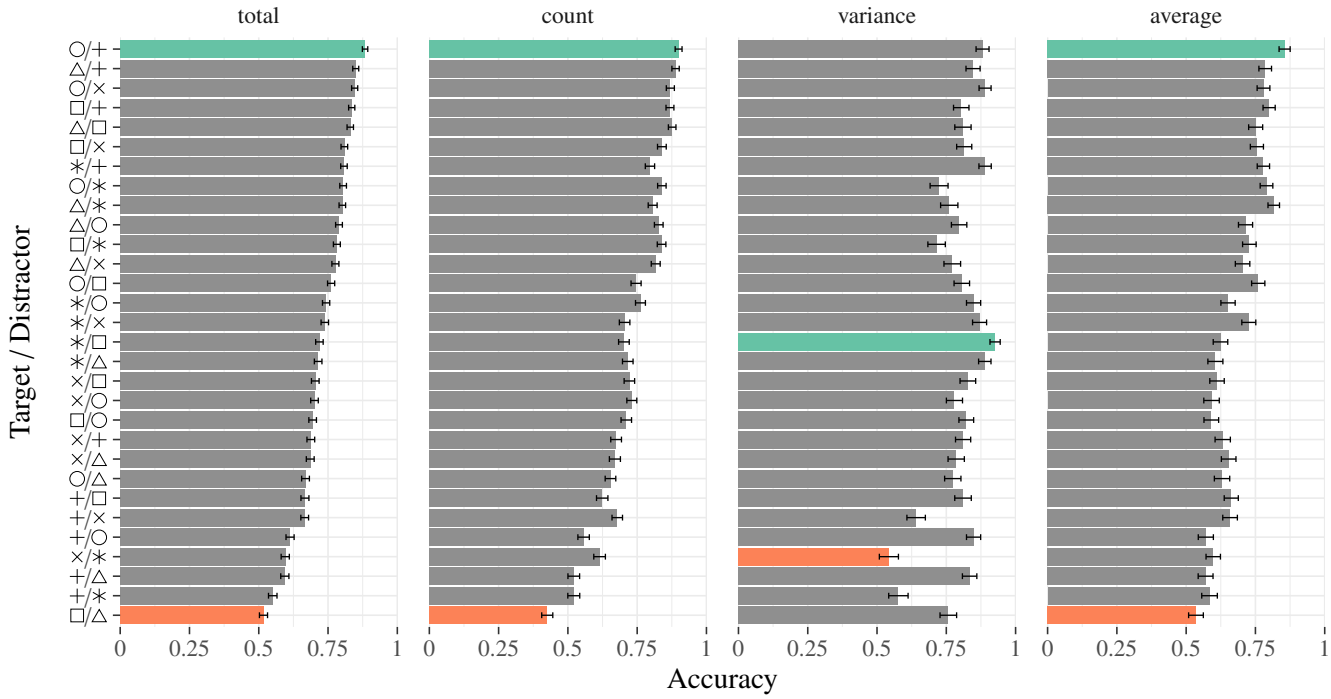

Fig. 11. Participants accuracy using different combinations of target and distractor shapes. Ordering is done based on accuracy over all experiments using our three tasks. The left barchart shows a comparison of accuracy using all tasks, while the remaining show accuracy using this individual tasks. Green bars show the combinations which showed the highest accuracy, while red bars show the combinations with the lowest accuracy.

through the Vuong's Closeness test [30]. Thus, for each of our tasks, we created a logistic regression model and evaluated the relative predictive performance. Furthermore, we computed the Akaike information criterion (AIC), to additionally argue about the relative quality of our models.

Comparing Number of Shapes. As described before, we first create a null model before subsequently adding variables of the scatterplot to this null model. For the comparing number task, these predictive variables are target shape, distracting shape, and number of shapes. For both experiments we conducted using the number of shapes task, we found that these variables could improve the predictive performances of the model significantly with $p < .001$ for each of the variables. Using this model we then applied our metric to investigate if our metrics can be used as predictive variables. We found that all our metrics could improve the predictive performances of the model significantly ($p < .001$) for both experiments. When adding our metrics to the model we found that for Experiment 1 (using the big shapes), all of the metrics could improve the model significantly ($p < .001$ for all models containing the different metrics). However, for Experiment 2 we found a significant increase in performances for M_{rel} ($p = .04551$), but not for M_{num} ($p = .07758$), M_{pix} ($p = .06235$), and M_{shape}($p = .6311$). Thus, for Experiment 1 all metrics could improve the model significantly, and also no significant difference could be found when comparing these models containing our metrics. This indicates that for small amounts of overlap M_{rel} could describe the data the best, while for larger amounts of overlap all metrics fit equally well.

Comparing Variance. Since the number of shapes is the same for all the scatterplots, we used target shape, distracting shape and variance as predictive variables for our base model of this task. For our experiment using the big shapes, target, as well as distracting shape, could improve the model significantly ($p < .001$). The same is true for the experiment using the small shapes, where target and distracting shape could again significantly improve the model ($p < .001$). For the experiment using the big shapes, however, using the variance as a predictive variable could not improve the model ($p = .797$) whereas for the small shapes using the variance could improve the model significantly ($p < .001$). So for further investigations, if our metrics could also serve as a predictive variable, we used a model with target shape, distractor shape and variance for the experiment using the small shapes, and a model containing target and distractor shape for the experiment using the big shapes. When further adding our metrics to these models, we found that again all the metrics could improve the model significantly ($p < .001$). We then compared all the models with each other using Vuong's Closeness test and found no significant difference between the models for the experiment using the big shapes. However, during analysis using AIC we found a better fit of the model containing the shape overlap metric (AIC : $M_{shape} = 2570$, $M_{rel} = 2577$, $M_{num} = 2579$, $M_{pix} = 2580$).

When comparing the models using AIC, we found that for this experiment, M_{pix} showed the best fit to the data (AIC : $M_{shape} = 3261$, $M_{num} = 3266$, $M_{rel} = 3275$, $M_{pix} = 3264$). By futher comparing the models using Vuong's Closeness Test, we

also found significant effects for this better fit (Experiment 4: $M_{shape} > M_{rel}$, $p = .0023479$).

Comparing Average. In our last task, the mean between the sets was varied, therefore we added the mean as a predictive variable besides target and distracting shape. For both of the experiments, we found that these three variables could improve the predictive performance of the model significantly ($p < .001$ for all variables in both experiments). We then again added our metrics to the models and found that for Experiment 5 M_{rel} ($p = .0411$), M_{pix} ($p = .01221$), and M_{shape} ($p = 0.004627$) could significantly improve the model, while M_{num} could not. When comparing the model containing M_{rel} with M_{pix}, we found no significant increase in performance, but a slight increase for the AIC criteria ($AIC : M_{shape} = 6477$, $M_{num} = 6484$, $M_{rel} = 6481$, $M_{pix} = 6479$). For Experiment 6, none of our metrics could improve the model significantly and the AIC values are almost equal ($AIC : M_{shape} = 5327$, $M_{num} = 5327$, $M_{rel} = 5327$, $M_{pix} = 5326$). We suppose this to happen, since the overall accuracy of participants in this experiment was rather low, which results in data that is difficult to predict for our regression model.

Combined Data from all Tasks. Since our goal is to find a model that works for different tasks commonly used in scatterplot analysis, we then tried if our metrics can also be used as a predictive variable for the complete data acquired in all our experiments. We suspected this to be possible, since all task parameters like amount of shapes, variance, shape sizes, and mean can be used as predictive variables. Therefore, we created models using all these variables and compared the models as described earlier in this section.

Target shape, distracting shape, amount of points, variance and mean showed a strong significant improvement to the model ($p < .001$ for each of the variables), while size of shape only showed a weak significant effect ($p = .04914$). When adding our metrics to the model we found that M_{shape} and M_{pix} could significantly improve the predictive performance of the model (M_{shape} $p < .001$; M_{num} $p = .385603$; M_{rel} $p = .373639$; M_{pix} $p = .026617$, $AIC : M_{shape} = 37250$, $M_{num} = 37270$, $M_{rel} = 37270$, $M_{pix} = 37260$), suggesting that this metric serves as the best predictor for human perception. Also when comparing $M_{shape}$and M_{pix}using Vuong's Closeness test, we found a significantly better fit in favour of M_{shape} $p = .0050261$.

9. Implications for Scatterplot Design

Even though the choice of target and distractor shape is important when designing scatterplots, our findings indicate that the overlap of shapes is also of great importance. Thus, while previous work could show that there are visual differences between shapes and combinations thereof [11, 20, 10, 22], we could show that overlap needs to be considered when transferring these findings to real world scatterplot scenarios. This is especially relevant since our findings indicate that there are combinations of shapes that suffer stronger from large amounts of overlap than others. Figure 12 shows this effect of overlap on the response accuracy for selected shapes. While there are combinations such as the circle (○) and plus (+) symbol which

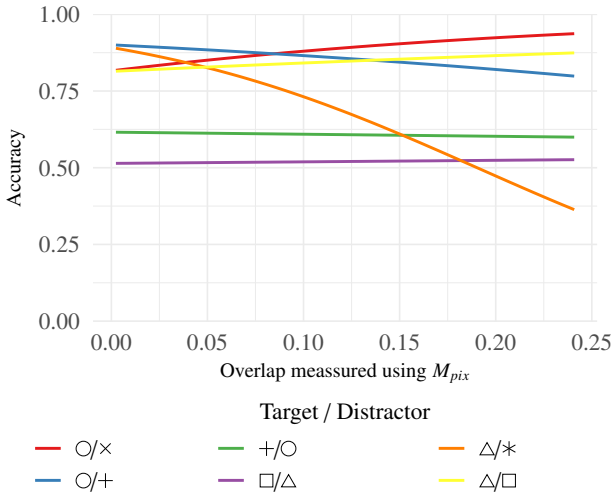

Fig. 12. Logistic regression curve of participants accuracy while overlap increases. This figure shows combined results taken from all our experiments using three different tasks and two different shape sizes for each task. The overlap was meassured using metric M_{pix} .

worked well for all tasks and overlap conditions, there are other combinations such as for instance the triangle (△) and the asterisk (∗), which seem to suffer severely from an increasing amount of overlap. This finding is especially relevant because both of these combinations combine closed target and open distractor shapes as suggested by Burlinson et al. [11]. Thus, while we could, in general, confirm that their findings of open vs. closed shapes also hold when incorporating overlap, our findings show that some combinations are still not beneficial to be used in practice.

Also, the given task seems to be an important factor when comparing conditions where shapes overlap. The usage of the asterisk (∗) shape as target seems to work well in combination with all other tested shapes as distractor for the variance task (see Figure 1), indicating that the saliency of the asterisk seems to be an important factor especially when trying to identify clusters. The combination of plus (+) and asterisk (∗) shows a similar effect for the number of shapes task. While this combination, in general, shows bad accuracies over all tasks (see Figure 11), it shows especially bad results for the number of shapes task and when overlap increases (see Figure 8). Investigating the specific benefits or cost of the asterisk (∗) shape remains future work.

Table 1 shows a comparison of all accuracies of all used shapes as well as all our used tasks. While this table indicates that the combination of closed target and open distractor shapes seems to work well in general, the combination of asterisk (∗) and plus (+) seems to be an interesting outlier, which works well for all tasks.

9.1. Predicting Perception of Scatterplots

Since our findings suggest that overlap of shapes is an important factor on human perception when perceiving scatterplots, the amount of overlap appearing in a given scatterplot should not be neglected when designing a scatterplot. To incorporate overlap in the design process, a logistic regression model could

be used to predict the readability of a scatterplot in the light of expected overlap. For this prediction to be accurate however, there is a need to determine all significant factors which influence the outcome of the prediction. We could show, that shape overlap is one of those factors and that when taking overlap into account, the predictive performance of a logistic regression model significantly improves. To find weight factors for such a predictive model, we would suggest using our pixel-based metric, which improves the predictive performance the most. Thus, by using these weight factors for overlap, used shapes, amount of data points, variance, and other factors, optimal parameters for a scatterplot visualization could be predicted through the resulting equation. Such a equation could look like the equation for the a logistic regression as follows:

$$p(x) = \frac{1}{e^{-(\beta_0+\beta_1+\ldots+\beta_y x)}} \tag{1}$$

Where β_x are weight factors for the scatterplot parameters and $p(x)$ is the possibility of an observer giving the correct answer to a given question. Using appropriately chosen predictive variables for β, the possibility of an intended answer by an observer can be maximized and hence the scatterplot optimized.

However, for an optimal predictive performance of the model, a larger amount of human labeled data is needed in order to find accurate weight factors for the regression model. Therefore, a larger number of different shape sizes, amounts of shapes, and further tasks might be needed. Also, when using more than two types of shapes, or even more complex tasks, additional parameters are introduced which also need to be investigated. This however again requires a larger amount of required data, since the number of combinations for these different parameters grows substantially. Thus, we see our findings as an essential ingredient for a predictive scatterplot model, but believe that more data is necessary to formulate such a model.

Findings. Even though we did limit our investigations to a two-way comparison by using two different shapes together at a time, we found some combination of shapes to appear to be less prone to overlap than others. Thus, based on our findings, we suggest for future multiclass scatterplot designs to use the following shape combinations: circle (○) and cross (×), circle (○) and plus (+), or triangle (△) and plus (+), as these combinations showed the overall best performance in terms of accuracy over all our experiments (see Figure 11). However, with an increasing amount of overlap and especially for tasks involving the perception of individual shapes, like in our number of shapes task, even these combinations suffer from overlap. We therefore further suggest to generally minimize overlap by choosing smaller shape sizes without jeopardizing shape readability. This suggestion is supported by our finding that when comparing different shape sizes, no significant effect of only shape size on participants' accuracy could be found.

10. Conclusion & Future Work

When exploring data using scatterplots, the ability to compare the given classes depends on the given data. Similarity, for example in distributions or increasing numbers of data points leads to an increasing degree of overlap which can obscure data. This leads to a need for optimization of the draw parameters of a given scatterplot to enable observers to explore the data. Therefore, within this paper, we presented the results of a series of crowd-sourced user studies that have been conducted to investigate the perceptual influence of overlap in two-dimensional scatterplots.

Table 1. Comparison of all combinations of target and distractor shapes, used in our experiments. A + symbol indicates that this combination resulted in 80% or more correct answers for the given task, while a – symbol indicates an accuracy of less than 60%. Combinations which showed high accuracy for all task are marked in green, while combinations with low accuracy are marked in red.

Target \ Dist.	□	○	△	+	×	∗	Task
□			–	+	+	+	Number
		+		+	+		Variance
		–	–				Average
○				+	+	+	Number
				+	+		Variance
				+			Average
△	+	+		+	+	+	Number
	+			+			Variance
						+	Average
+		–	–			–	Number
	+	+	+			–	Variance
		–	–			–	Average
×							Number
	+			+		–	Variance
		–				–	Average
∗							Number
	+	+	+	+	+		Variance
							Average

While research suggests color as the most dominant visual channel, recent work could show that using shapes to encode can be a viable choice as well. In contrast to color, where drawing-order is the most important factor in the ability to perceive datapoints, the overlap between shapes introduces different visual artifacts like artificial new shapes. Therefore, we found that overlapping shapes can have a strong influence on human perception of scatterplots and therefore need to be taken into account. o measure this overlap we presented three different metrics and compared them using a logistic regression model. While we found that in some cases even simple measurements can be used as a predictive variable, our pixel precise metric showed the overall best predictive performance.

While we could show that, in general, overlap influences the ease of perception of scatterplots, we could also show that some shape combinations are less prone to an increasing amount of overlap. Furthermore, we could confirm previous work that closed shapes show a good distinctiveness, especially when used as target shape in combination with an open distractor shape [11]. However, the asterisk shape is an interesting out-

lier to this rule, when used as a target shape for example in the variance task.

Using our metrics and having representative, human labeled examples of scatterplots to be optimized, an automated approach could be implemented to enhance the distinctiveness of the plot and hence improve human perception. While our evaluations are limited since we only used two different types of shapes at the same time, and a rather small amount of data points, we could still show that the pixel-based metric could serve as a valuable predictive variable for common tasks when investigating scatterplots.

In the future, we would like to investigate if our metric can be used to predict participants' performance for more complex scatterplots. These complex scatterplots could involve more than two different shapes at a time in a single scatterplot. We suspect that there is an additional interaction between certain combinations of shape (for example when using shapes which form another shape which also occures in the plot (for example cross ($\times$), plus ($+$), and asterisk ($*$)).

Using different colors also introduces additional complexity, since the order of overdraw becomes relevant when compared to drawing using only one color. Furthermore, we would like to investigate if our findings can be extended to a larger amount of data points, and prove if our findings can be interpolated for example for different sizes of shapes between seven and fifteen as used in our experiments.

References

[1] Matejka, J, Anderson, F, Fitzmaurice, G. Dynamic opacity optimization for scatter plots. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM; 2015, p. 2707–2710.

[2] Micallef, L, Palmas, G, Oulasvirta, A, Weinkauf, T. Towards perceptual optimization of the visual design of scatterplots. IEEE Transactions on Visualization and Computer Graphics 2017;23(6):1588–1599.

[3] Sarikaya, A, Gleicher, M. Scatterplots: Tasks, data, and designs. IEEE Transactions on Visualization and Computer Graphics 2017;.

[4] Bachthaler, S, Weiskopf, D. Continuous scatterplots. IEEE Transactions on Visualization and Computer Graphics 2008;14(6):1428–1435.

[5] Correll, M, Heer, J. Regression by eye: Estimating trends in bivariate visualizations. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. ACM; 2017, p. 1387–1396.

[6] Chen, H, Engle, S, Joshi, A, Ragan, ED, Yuksel, BF, Harrison, L. Using animation to alleviate overdraw in multiclass scatterplot matrices. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. ACM; 2018, p. 417.

[7] Christ, RE. Review and analysis of color coding research for visual displays. Human factors 1975;17(6):542–570.

[8] Mackinlay, J. Automating the design of graphical presentations of relational information. Acm Transactions On Graphics (Tog) 1986;5(2):110–141.

[9] Lewandowsky, S, Spence, I. Discriminating strata in scatterplots. Journal of the American Statistical Association 1989;84(407):682–688.

[10] Li, J, van Wijk, JJ, Martens, JB. Evaluation of symbol contrast in scatterplots. In: Visualization Symposium, 2009. PacificVis' 09. IEEE Pacific. IEEE; 2009, p. 97–104.

[11] Burlinson, D, Subramanian, K, Goolkasian, P. Open vs. closed shapes: New perceptual categories? IEEE Transactions on Visualization and Computer Graphics 2017;.

[12] Li, J, Martens, JB, van Wijk, JJ. A model of symbol size discrimination in scatterplots. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM; 2010, p. 2553–2562.

[13] Szafir, DA. Modeling color difference for visualization design. IEEE Transactions on Visualization and Computer Graphics 2017;.

[14] Sedlmair, M, Aupetit, M. Data-driven evaluation of visual quality measures. In: Computer Graphics Forum; vol. 34. Wiley Online Library; 2015, p. 201–210.

[15] Sips, M, Neubert, B, Lewis, JP, Hanrahan, P. Selecting good views of high-dimensional data using class consistency. In: Computer Graphics Forum; vol. 28. Wiley Online Library; 2009, p. 831–838.

[16] Tatu, A, Albuquerque, G, Eisemann, M, Schneidewind, J, Theisel, H, Magnork, M, et al. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In: Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on. IEEE; 2009, p. 59–66.

[17] Ross, J, Irani, L, Silberman, M, Zaldivar, A, Tomlinson, B. Who are the crowdworkers?: shifting demographics in mechanical turk. In: CHI'10 extended abstracts on Human factors in computing systems. ACM; 2010, p. 2863–2872.

[18] Buhrmester, M, Kwang, T, Gosling, SD. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? Perspectives on psychological science 2011;6(1):3–5.

[19] Heer, J, Bostock, M. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM; 2010, p. 203–212.

[20] Demiralp, Ç, Bernstein, MS, Heer, J. Learning perceptual kernels for visualization design. IEEE transactions on visualization and computer graphics 2014;20(12):1933–1942.

[21] Tremmel, L. The visual separability of plotting symbols in scatterplots. Journal of Computational and Graphical Statistics 1995;4(2):101–112.

[22] Gleicher, M, Correll, M, Nothelfer, C, Franconeri, S. Perception of average value in multiclass scatterplots. IEEE transactions on visualization and computer graphics 2013;19(12):2316–2325.

[23] Mayorga, A, Gleicher, M. Splatterplots: Overcoming overdraw in scatter plots. IEEE transactions on visualization and computer graphics 2013;19(9):1526–1538.

[24] Keim, DA, Hao, MC, Dayal, U, Janetzko, H, Bak, P. Generalized scatter plots. Information Visualization 2010;9(4):301–311.

[25] Urribarri, DK, Castro, SM. Prediction of data visibility in two-dimensional scatterplots. Information Visualization 2017;16(2):113–125.

[26] Gilbert, EG, Johnson, DW, Keerthi, SS. A fast procedure for computing the distance between complex objects in three-dimensional space. IEEE Journal on Robotics and Automation 1988;4(2):193–203.

[27] Rensink, RA, Baldridge, G. The perception of correlation in scatterplots. In: Computer Graphics Forum; vol. 29. Wiley Online Library; 2010, p. 1203–1210.

[28] Bostock, M, Ogievetsky, V, Heer, J. D^3 data-driven documents. IEEE transactions on visualization and computer graphics 2011;17(12):2301–2309.

[29] Lagae, A, Dutré, P. A comparison of methods for generating poisson disk distributions. In: Computer Graphics Forum; vol. 27. Wiley Online Library; 2008, p. 114–129.

[30] Vuong, QH. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica: Journal of the Econometric Society 1989;:307–333.