The Influence of Contour on Similarity Perception of Star Glyphs

Johannes Fuchs, Petra Isenberg, Anastasia Bezerianos, Fabian Fischer, and Enrico Bertini

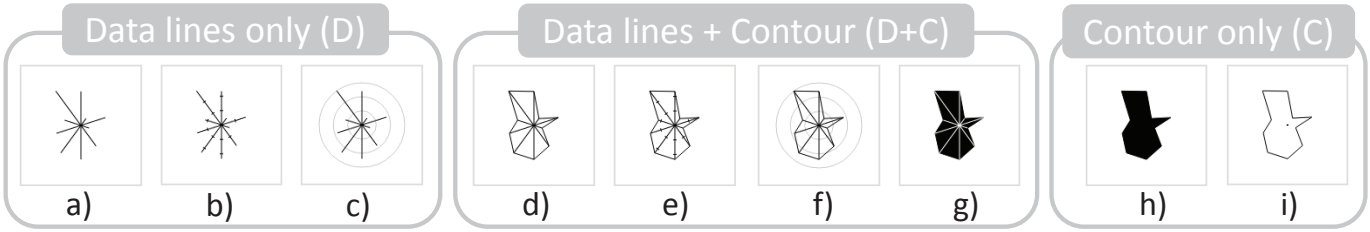

Fig. 1. Three categories of star glyph variations used in our three experiments. *Data lines only* (D): only the data lines encode the data; *Data lines+Contour* ($D+C$): data lines are connected at the endpoints to create a closed shape; *Contour only* (C): only the contour line is drawn. Additional variations are tickmarks (T) (in b, e)), gridlines (G) (in c, f)), and fill style (in g, h)).

Abstract— We conducted three experiments to investigate the effects of contours on the detection of data similarity with star glyph variations. A star glyph is a small, compact, data graphic that represents a multi-dimensional data point. Star glyphs are often used in small-multiple settings, to represent data points in tables, on maps, or as overlays on other types of data graphics. In these settings, an important task is the visual comparison of the data points encoded in the star glyph, for example to find other similar data points or outliers. We hypothesized that for data comparisons, the overall shape of a star glyph—enhanced through contour lines— would aid the viewer in making accurate similarity judgments. To test this hypothesis, we conducted three experiments. In our first experiment, we explored how the use of contours influenced how visualization experts and trained novices chose glyphs with similar data values. Our results showed that glyphs without contours make the detection of data similarity easier. Given these results, we conducted a second study to understand intuitive notions of similarity. Star glyphs without contours most intuitively supported the detection of data similarity. In a third experiment, we tested the effect of star glyph reference structures (i.e., tickmarks and gridlines) on the detection of similarity. Surprisingly, our results show that adding reference structures does improve the correctness of similarity judgments for star glyphs with contours, but not for the standard star glyph. As a result of these experiments, we conclude that the simple star glyph without contours performs best under several criteria, reinforcing its practice and popularity in the literature. Contours seem to enhance the detection of other types of similarity, e. g., shape similarity and are distracting when data similarity has to be judged. Based on these findings we provide design considerations regarding the use of contours and reference structures on star glyphs.

Index Terms—Glyphs, star glyphs, contours, perception, quantitative evaluation, similarity detection, visual comparison

1 Introduction

Data glyphs are small composite visual representations of multi-dimensional data points. Glyphs express the dimensions of a data point by assigning them to a specific visual variable [34]. Given their small graphical footprint glyphs are very versatile, used in a variety of different application areas: Monitoring computer networks [11, 16], tracking the health of patients [22], comparing country characteristics [23], or analyzing sports games [21]. Glyphs, in contrast to general charts or other visualizations, are often used as small visual representations nested inside other visualizations such as hierarchies, networks, or geographic data—or when a very large number of data points needs to be seen in one overview. Their primarily role is typically to provide quick overviews and help detect data trends and similarities [34].

A star glyph [29] is a specific type of glyph that lays out the axes for each data dimension on a radial grid and matching the dimension's values to a position on the respective axes, typically connected with a line to the center of the glyph. There exists a great variety of alternative designs for star glyphs that differ in the amount of reference structures used, the use of additional visual variables on the "rays," or whether or not the individual rays are connected to form a contour for the glyph [33]. The version of the star glyph with unconnected rays is also sometimes called *whisker or fan plot*, while the connected version also carries the name *star plot* [34]. Star glyphs are frequently used but very little advice exists on how to choose between different star glyph encodings. The question arises to what degree changes in the design of a star glyph influence its perception and, thus, the effectiveness of the glyph in certain tasks.

One important task for glyphs in small-multiple settings is the comparison of the encoded data points to one-another. Such a comparison task may be conducted to find data points that are very close over all dimensions, very different, or similar in just a subset of dimensions. We focus on the first task: finding data points encoded as star glyphs that are very similar to a target glyph. We are interested in this task because if it is well supported, it should improve people's ability to perform the other two types of comparison tasks. We hypothesized that the ability to perceive a star glyph as a coherent and closed shape would strongly influence the correctness of data similarity detection tasks—as it would potentially be easier to compare a single shape than having to compare individual rays. This hypothesis was motivated by prior research showing that a closed contour has an influence on the perception of a coherent shape [8]. As Palmer noted: *"Shape allows a perceiver to predict more facts about an object than any other property"* [24].

We are not aware of any previous studies on the importance of glyph contour on tasks with multi-dimensional data. We contribute three studies, showing the differences in performance when adding a contour structure to well-known glyph designs. In the first we investigate the effect of contours on the perception of data similarity with data

- *Johannes Fuchs is with the University of Konstanz. E-mail: fuchs@dbvis.inf.uni-konstanz.de.*
- *Petra Isenberg is with Inria. E-mail: petra.isenberg@inria.fr.*
- *Anastasia Bezerianos is with Université Paris-Sud, CNRS & Inria. E-mail: anastasia.bezerianos@lri.fr.*
- *Fabian Fischer is with the University of Konstanz. E-mail: fabian.fischer@uni-konstanz.de.*
- *Enrico Bertini is with NYU Poly. E-mail: ebertini@poly.edu.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014. Date of publication 11 Aug. 2014; date of current version 9 Nov. 2014. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.
Digital Object Identifier 10.1109/TVCG.2014.2346426

visualization experts and novices on star glyphs. Our results indicate that contours influenced expert the most, whereas novices had poor performance across the board. Nevertheless, we observed that some types of shape-related similarities overpowered the participants' ability to read data. Based on this result we conducted a second study on the nature of similarity perception for a larger set of glyphs with and without contours. We found that, even without any participant training, glyphs without contours led to similarity judgments that are close to data reading, thus making them better candidates for visualization. In our third and final study, we added simple reference structures—tick marks and grids—to the designs, and examine how they affect similarity judgments. Our results show that reference structures aid data similarity comparisons in star glyphs with contours, but have little effect on ones without. Based on these results we present considerations for the design of more effective glyphs for similarity detection tasks.

2 Motivation

There are many real-world scenarios where multi-dimensional glyphs can provide valuable information. Multi-dimensional data is notoriously hard to represent visually as the number of visual variables available to encode data dimensions is limited. Multi-dimensional glyphs, and more specifically glyphs where data dimensions are presented through radial axes, provide "hints" of the underlying multi-dimensional structure when multi-dimensional objects are plotted on the spatial substrate. Common examples are: maps showing the geographical distribution of multi-dimensional objects (e. g., comparison of indicators such as crime rate or suicides for different regions of France [12]), multi-dimensional scaling visualizations exposing relationships between scaling algorithms and data distributions (e. g., election patterns to show political party proportions by region [30]), or data objects organized in a grid layout to show how multi-dimensional objects distribute across sets of predefined categories (e. g., food nutrients in different food categories).

In all these examples glyph similarity plays a prominent role: "*Which regions have similar political characteristics? Do all the glyphs in this cluster have similar data values? Do fruits and vegetables have similar nutrients?*" Many of these kinds of questions require that glyphs represent similarity across data points accurately.

Our studies show that the accuracy with which this similarity is perceived is not equal for all glyph designs and that some design solutions can lead to more truthful estimations. Being able to gauge which elements of glyph design may lead to more accurate estimations is crucial. More accurate designs may lead to at least two major improvements: increase in confidence with which end-users estimate data similarity and ultimately more accuracy in data interpretation.

3 Related Work

For a detailed overview of research on data glyphs, we refer the interested reader to two summary articles [4, 32]. Our work, in contrast, focuses on one specific glyph—the star glyph. In this respect, we were inspired by two main bodies of work: a) related studies investigating the performance of star glyphs in information visualization for multi-dimensional data, and b) studies focusing on the perception of two-dimensional shapes.

3.1 Evaluation of Star Glyphs

Similar to our work, several researchers have studied the perception of glyphs in the context of similarity tasks—yet with a variety of methodological approaches. Wilkinson [35] conducted a user study comparing star glyphs, castles, Chernoff faces and blobs. Participants had to sort 8 glyphs of each type—varied by a variety of factors—according to increasing dissimilarity. Their findings indicate that judgments on Chernoff faces were closer to the actual factor distances, followed by star glyphs, castles and blobs.

A similar sorting-based task was used by Borg and Staufenbiel [3] in their comparison of snowflakes (similar to star glyphs), suns, and factorial suns. Participants had to sort 3×44 shuffled cards showing data points of one type of glyph into four categories according to their similarity. Factorial suns—that make use of some preprocessing of the data—were most easily discriminated and star glyph performed the worst in this respect.

Lee et al. [20] showed participants several datasets represented by one of: small-multiples Chernoff faces, star glyphs, and two plots produced with multi-dimensional scaling. For each dataset participants were given eight questions to answer, some of which included similarity judgments based on pairwise comparisons. The authors did not perform an analysis on the basis of individual similarity questions. Instead, they found that participants performed best and were most confident with one of the 2D spatial plots, in particular on global questions where the whole set of data points has to be considered.

Klippel et al.'s study [17] is probably the most related to our work as it also studied the influence of shape on glyph perception based on similarity judgments. Yet, instead of the influence of contour, as in our case, they varied shape by reordering the dimensions in a star glyph with contour. The authors studied how shape changes influenced the interpretation of data points in a similarity-based grouping task. They found that differences in shape influenced cognitive processing of the data and that perceptually salient features (such as spikes) strongly influenced how people thought about a data point.

In summary, a wide variety of approaches have been taken to study glyph similarity and these vary in methodology and the factors of the glyph designs. While our study differs in methodology and the factors studied from this previous work, we share the question of how design factors influence the perception of data similarity (Δ of changes to all data values) vs. shape similarity (perceived similarity of shape).

3.2 Perception of Two-Dimensional Shapes

A number of experiments in the perceptual psychology literature have investigated the effect of contour closure on shape perception. Contour closure is a phenomenon related to the Gestalt principle of closure [18] according to which we perceive visual objects as grouped together if they seem to complete a visual entity. Visual objects can, thus, have an "open contour" of unconnected marks (dots, lines, etc.) or a completely "closed contour" that forms one continuous boundary. Several researchers have tested open vs. closed contours and found that closed contours were perceptually superior to open contours in a variety of different tasks:

Elder and Zucker [8] showed that the efficiency of visual search was better for shapes with a closed contour. Similar to our studies, participants were shown a stimulus object, which had to be found amongst a larger set of distracters. In later work [9], the authors found supporting evidence that geometric region boundaries, such as contours, are processed much earlier by our perceptual system than other surface features, such as a region's texture.

Garrigan [14] investigated the recognition accuracy of stimuli with a closed contour compared to those with an open contour. Participants had to learn a set of stimuli and later recall whether a newly presented stimulus had been in the previously learned set of objects or not. The experiment showed that the closed contour shapes were recognized more easily. The authors argue that this is due to a better encoding of the stimuli in the human visual system. Finally, Saarinen et al. [28] measured the precision of shape perception. Participants had to judge whether the aspect ratio of the stimulus (i. e., a rectangle) was tall or wide. Their results showed that for discrimination tasks, closed contour shapes were more effective than non-closed shapes.

In summary, a large body of literature exists on the effect of contours on shape perception. We highlighted a few that have particularly inspired our hypotheses that the presence of contours may be of particular importance to the effective use of glyphs in visual data analysis tasks. We contribute a further study focused on a concrete application in visual data analysis, with glyphs showing multi-dimensional data points. We examine in particular whether glyph shape, emphasized by a closed contour, is an important predictor for the effectiveness of glyph designs in a similarity detection task.

4 Experiment 1: Contours for Novices vs. Experts

In our first study we were interested in the fundamental question: does contour affect people's perception of data similarity with star glyphs?

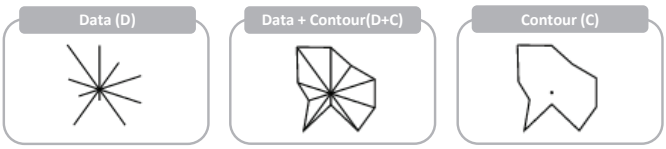

Fig. 2. **Experiment 1 Contour Variations**: (from left to right) star glyph with rays and no contour (D); common star glyph ($D+C$); only the contour line of the star glyph (C).

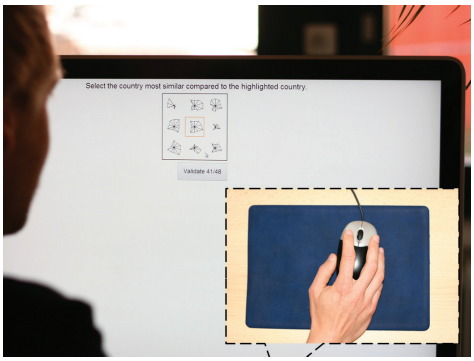

Fig. 3. **Experiment Setting**: The participant was seated in front of a 24" Screen with a resolution of 1920x1200. The only input device was a computer mouse.

Data similarity judgments are cognitive tasks, where the viewer has to judge the absolute difference in all dimension data values between two data points. This differs from other types of similarity judgments, such as detecting shape similarity e. g., under rotation or scale.

Detection of data similarity is a synoptic task according to the Andrienko & Andrienko [1] task taxonomy. Synoptic tasks are very common and important for glyphs in small-multiple settings. Analysts have to visually compare data points to detect outliers or to identify similar groups of data points, by referring to the whole data set or a subset of the data (e. g., finding countries with similar characteristics).

We were interested in the effect of contour, as we hypothesized—based on previous perception studies [8]—that a contour would impact the rapid perception of shapes and, thus, aid in tasks that require the data point to be perceived in its entirety. Finally, we hypothesized that there would be a difference between experts' and novices' ability to make accurate data similarity judgments, and thus chose to conduct a between-subjects experiment with these two groups of participants.

4.1 Design and Procedure

Glyphs: We used three different variations of the star glyph (Fig. 2). The first, also called whiskers or fan plot [26, 34] uses "rays" to encode quantitative values for each dimension through the length of each ray. We refer to this variation as "Data lines only (D)". The second variation, "Data lines + Contour ($D+C$)", connects the end of each ray with a line to add a closed contour [29]. In the third variation the radial rays are removed, and only the contour line is presented [6]. We use the term "Contour only (C)" for this design variant. All three star glyph contour variations have been used in real-world context and in the scientific literature, thus adding external validity of our glyph choice.

Dimensionality: To investigate the effect of contours on different data densities we varied the number of dimensions shown in the glyphs. The *low* dimension density consisted of three data dimensions with corresponding data values, while the *high* density consisted of ten data dimensions. We considered ten dimensions to be *high*, as glyphs used in the literature rarely visualize more than ten dimensions; also to our knowledge there is no study investigating the maximum number of perceivable dimensions in a single star glyph to use as a basis.

Task, Procedure and Apparatus: Participants were shown a highlighted stimulus glyph surrounded by 8 more glyphs in a 3 × 3 matrix configuration (Fig. 3). One of these glyphs was closest in data space (lowest absolute data distance) while the rest were distracters further away in data space. The participant had to select the glyph closest to the stimulus in terms of data value. For each contour variation, participants were given training explaining how data was encoded and the notion of similarity in data space. They were then given four practice trials where the correct answer was revealed to help learning. During the actual experiment the correct answer was no longer provided.

The three glyph variations were presented in an order randomized using a latin square. The position of the correct answer as well as the different distracters was also randomized. Similarly, the exact glyph values were randomized (as discussed in Section 4.3). Each participant repeated 4 training and 4 real trials for each contour variation.

The study took place in a lab setting in the presence of an experimenter. The experiment was conducted on a 24 inch screen with a resolution of 1920 ×1200 and took around 25 minutes. The only input device was a common computer mouse to make selections.

Participants: Twelve novices (7 female) and twelve experts (2 female) participated in our study. The age of novice participants ranged from 18–23 years (mean & median age 20), and from 26–38 years (mean 30.3 and median 29) for experts. All participants reported normal or corrected-to-normal vision. All novice participants reported no experience in reading glyphs, but were familiar with common chart visualizations seen in print (e. g., bar and pie charts). All 12 experts were visualization researchers and students who reported a strong background in data visualization with at least basic knowledge of reading glyphs (1 Bachelor; 8 Master; 3 PhD).

Overall our experiment consisted of:

3	contour variations (D,$D+C$,C)	×
2	dimensionalities (*high*, *low*)	×
4	repetitions	=
24	trials per participant	×
24	participants (12 per expertise)	=
576	**trials in total**	

Expertise was the between-subjects factor.

4.2 Hypotheses

H1: Novices are less accurate in judging data similarity than experts.
H2: Both experts and novices make more accurate judgments in the low dimensional than the high dimensional condition.
H3: For both experts and novices, contour variations ($D+C$,C) improve the accuracy of data similarity judgments.
H4: This effect will be stronger in novices who have no prior glyph reading experience.
H5: Contour variations ($D+C$,C) lead to more accurate judgments mostly in the high dimensional condition, while the low dimensional condition is be less affected overall.

4.3 Data Generation and Distracters

Our data was synthetically created: 3 dimension values for the low and 10 for the high dimensional case. For each dimension we consider data values ranging from 0 to 5, partitioned in three value categories: low [0,1], middle [2, 3], high [4, 5]. We avoided larger value ranges as we were not interested in studying visual acuity.

The stimulus (i. e., central highlighted glyph) was created randomly by assigning either a middle or a high data value to the different dimensions with an equal chance of 25% (i. e., 50% for each value categories and 50% for the final data value). This was done once for all repetitions. To avoid learning effects, the stimulus was rotated between repetitions, keeping the values and the neighboring dimensions identical.

Each trial also contained a *target* glyph—the correct answer, thus the most similar to the stimulus in terms of data closeness (minimum data value distance). To generate it, we changed the data values of the stimulus randomly up to a maximum of 7 changes in data distance for the high dimensional condition, and 1 for the low. This was done by sequentially scanning the dimensions with a probabilistic function,

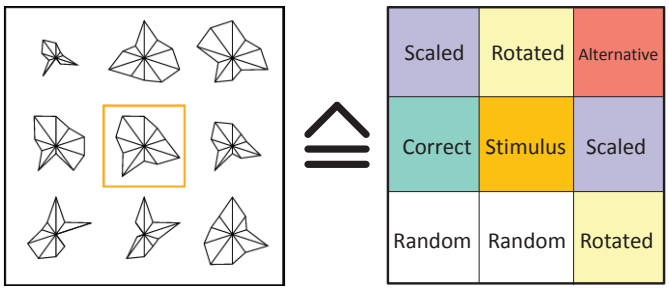

Fig. 4. **Experiment Setting**: For each trial glyphs were arranged in a 3×3 matrix. The stimulus is highlighted and positioned in the middle to assure an equal distance to the other glyphs. This setting is used in all three experiments.

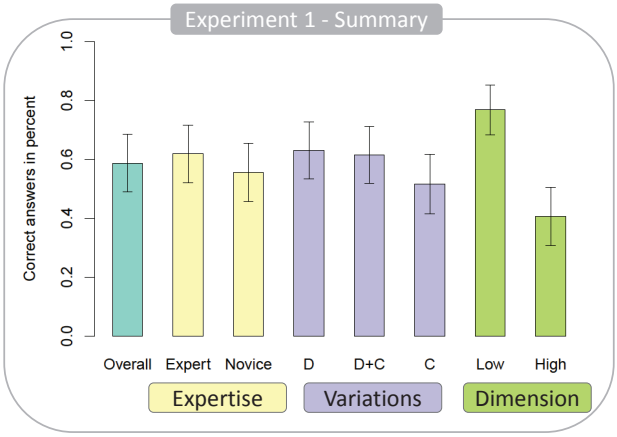

Fig. 5. **Experiment 1 Summary**: The bar charts illustrate the percentage of correct answers and the standard deviation.

which first decided to change the dimension or not (50%), second to increase or decrease the corresponding data value (50%) and third by how much (i.e., 1 or 2)(50%). At the end we ensured that the resulting data values fit into one of the three categories (i. e. low, middle, and high) and that the sum of all changes meet the predefined criteria.

Besides the stimulus and target glyph, we created 3 types of *distracters*. First, a *rotated* version of the stimulus, keeping the data values identical, but switching the dimensions one step either to the left or to the right. Second, a *scaled* version of the stimulus where we reduced the data values of each dimension by 1. Since the data values of the stimulus reach from 2 to 5 it is not possible to end up with negative values. Third, a close *alternative* of the target glyph. This alternative takes the data values from the stimulus and changes the values randomly up to a maximum of 8 changes in data distance for the high dimensional case, or 3 for the low. Values were chosen to ensure that the alternative glyph is not too different from the stimulus, while the target glyph continues to be the most similar in data distance. The remaining distracters were created randomly by assigning a data value to each dimension with an equal chance (see Fig. 4). For each trial we ensured that the sum of all differences between stimulus and all distracters was higher to that between stimulus and target glyph.

4.4 Results

We report only statistically significant results ($p < .05$) for accuracy. Given the non-normal nature of our data we used a non-parametric Friedman's test for the analysis of correct answers between glyph variations and a Kruskal-Wallis test for comparisons between expertise (between group factor). Fig. 5 shows overall correct answers, and Fig. 6 which type of distracters participants chose under the different experimental conditions. Although completion time was logged, we found no differences across variations and user groups, with low dimension trials taking on average $11sec$ ($D = 12.7sec$, $D+C = 11.3sec$, $C = 9,7sec$) and high ones $18sec$ ($D = 19.7sec$, $D+C = 16.9sec$, $C = 16.7sec$).

Overall accuracy for experts across variations was 79.1% for the low and 44.4% for the high dimensional glyphs, and for novices 74.3% and 36.8% respectively. However, there was no significant effect of *expertise* on *accuracy*. Fig. 5 illustrates more high level results.

Dimensionality: There was a significant effect of *dimensionality* on *accuracy* ($\chi^2(1,N = 288) = 23, p < .001$).

Post-hoc tests revealed that participants were more accurate in the low dimensional condition (76.7%) compared to the high dimensional condition (40.6%, $p < .001$).

Contour variation: There was a significant effect of *contour variation* on *accuracy* ($\chi^2(2,N = 192) = 7.9, p < .05$). Participants using variation C performed significantly worse (51.6%) compared to D (63%, $p < .05$) and $D+C$ (61.5%, $p < .05$). For experts, there was a significant effect of *contour variation* on *accuracy* in the high dimensional condition ($\chi^2(2,N = 48) = 12, p < .001$). A pairwise comparison revealed a significant higher accuracy with the D variation (66.7%) compared to both $D+C$ (41.7%, $p < .05$) and C (25%, $p < .001$). No significant results were found for novice participants.

When comparing the accuracy of the two participant groups, we found that for the variation D, there was a significant effect of *expertise* on *accuracy* in the high dimensional condition ($\chi^2(1,N = 96) = 5.85, p < .05$). Experts performed significantly better (66.7%) using the D variation compared to novice participants (39.6%, $p < .05$).

When selecting a wrong answer, both experts and novices most frequently selected the second closest data point to the stimulus (17.7%, 20.5% respectively), followed by a scaled version of the stimulus (16%, 16.3%) and to a lesser extent rotated versions (2.4%, 4.1%), mostly in the high dimension case of the contour variations ($C+D, C$).

4.5 Summary and Discussion

Overall we cannot confirm H1, our experts were not significantly more correct than novices on average. This is especially true for the low dimensional condition where both user groups had a good performance ($\approx 80\%$ correct). However, for higher dimensionalities experts using variation D were significantly more accurate compared to novices (partially confirming H1).

When comparing the two dimensionalities, similarity judgments were significantly more accurate for both user groups in the low dimensional condition compared to higher dimensionalities, confirming H2. With an increasing number of dimensions more data values have to be visually compared, leading to more complex mental calculations resulting in a higher error rate.

Contrary to intuition from previous work that contour can improve similarity judgments [8,14], we found that contour affected the accuracy of judgments negatively for experts. Thus we cannot confirm H3. As no significant effects were found for novice participants, we could also not confirm H4, however, mean accuracy for C (50%) was lower compared to $D+C$ (59.4%) and D (57.3%).

We also could not confirm H5. Contrary to expectations, the variation without a contour (D) led to significantly more correct answers for high-dimensional glyphs. The effect was not visible in the low dimensionality case where all participants were overall approx. 80% accurate with all variations.

Trying to explain the unexpected negative effect of contour on experts, especially in high dimensional cases, we noted that at least half of the erroneous answers in the contour variations ($C+D, C$) were in the form of scaled versions of the stimulus glyph, and to a lesser extent rotated versions, i. e. glyphs that have a geometric form similar to the stimulus glyph. In retrospect, this negative effect of contour could be explained by the fact that contour, and closure in general, is one of the factors promoting the notion of unity according to Gestalt psychology [18]. In our case contours led our experts to erroneously consider glyphs as coherent shapes when judging similarity, rather than data points. This resulted in judgments and comparison of geometrical shapes rather than data, with experts being led to consider as more similar data points that were either scaled or rotated versions of the stimulus, rather than the one closest in data space.

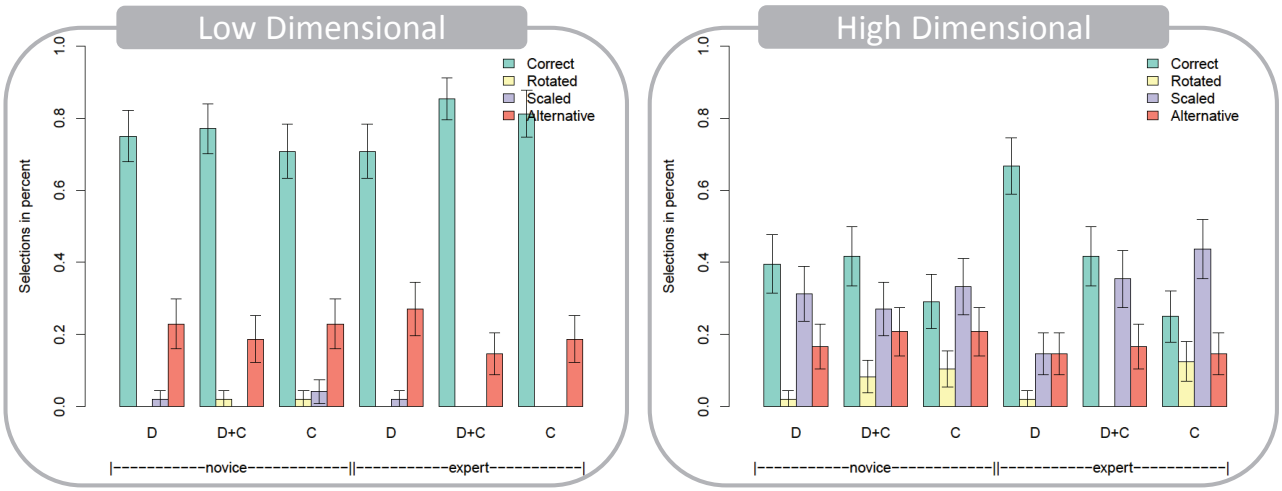

Fig. 6. **Experiment 1 Results**: The bar charts illustrate the percentage of selections and the standard deviation for each factor. In the high dimensional condition experts using variations $D+C, C$ were lead to judge shape similarity rather than data similarity whereas the accuracy of novices was low for all three variations.

Given the overall poor performance of novices in the high dimensional case we conjecture that due to their lack of familiarity and experience they tended to fall back to judging shape rather than data similarity for all star glyph variations. This is evidenced by the fact that at least half of their errors were a combination of scaled and rotated versions of the stimulus glyph.

5 Experiment 2: Perception of Similarity

Results from Experiment 1 indicated that in high dimensional cases contours mislead even experts to perceive rotated or scaled versions of the stimulus as more similar, rather than the one closest in data space. Based on this finding, we conducted a second experiment to better understand what type of similarity star glyphs naturally support. To this end, participants were not given any training or explanation of what similarity means, and we did not inform them that the glyphs encoded multi-dimensional data. Their only instruction was to select the most similar glyph. Our goal in this experiment was to examine what viewers naturally perceive as similar in different star glyph variations, without being instructed on how to judge similarity. Based on our results we hoped to identify the star glyph variations, if any, that naturally promote data similarity rather than shape similarity and, therefore, are more suitable for data visualization.

5.1 Design and Procedure

Glyphs. The experiment tested the glyph variations from Experiment 1, as well as a filled version of the C and $D+C$ glyph. We wanted to examine whether variations of glyphs that are filled reinforce more strongly the notion of a closed shape, due to the strong foreground/background contrast [18]. We conjectured that fill color may lead to more shape rather than data similarity choices. The experiment was a between-subjects design with fill type as the between-subjects factor. Thus, the D glyph was included in each group as the baseline. We had a total of 2 fill types (Fill, No-Fill) with 3 glyph variations each, as illustrated in Fig. 7.

Task. We again used a synoptic task, where participants selected the most similar glyph compared to a stimulus glyph. Participants were shown a highlighted stimulus surrounded by another 8 glyphs in a 3×3 matrix configuration. The positions of the surrounding glyphs were randomized around the stimulus. Again, we wanted to explore the notion of similarity and examine if some glyphs are naturally judged in a manner that approaches data rather than shape comparison. We thus gave no explanation as to what the glyphs represented and provided our participants with no training. Participants were free to interpret the word “similar” as they saw fit.

Data, Target Types and Dimensionality. Our data was generated as in Experiment 1, and again we tested *low* and *high* dimensionality. However, we included slightly different glyph choices to our participants, that we call "Target Types" (they are no longer distracters, as there is no correct answer). To balance the selection likelihood between each target type, we included two of each shape similarity and two glyphs that were closest to the stimulus in data space (we refer to this kind of target as “data"). As a result we had 2 data, 2 rotated and 2 scaled versions of the stimulus, and 2 randomly generated targets.

Participants and Procedure. Our study was conducted on Amazon Mechanical Turk (AMT), inspired by previous graphical perception experiments [5, 15]. We accepted 62 participants in total, and subjects were paid 0.50$ per Human Intelligence Task (HIT). Given the simple nature of our perceptual study, no qualification tests were required to complete our HITs. In accordance with AMT guidelines, however, only workers with 95% or more HIT approval rate were allowed to participate. Furthermore, we added control questions (3 in total) throughout the study, where one of the targets was identical to the stimulus and the answer was, therefore, obvious. We dismissed workers who did not get all the control questions correctly and their data was not included in the analysis. As a result we ended up with 36 participants (18 per fill type). Each participant worked on 4 trials for each variation and dimensionality, and viewed either the fill or the no-fill types. The order of presenting the glyph variations was randomized.

Overall our experiment consisted of

2	filling types (*Fill, No-Fill*)	×
3	contour variations (*D, D+C, C*)	×
2	dimensionalities (*high, low*)	×
4	repetitions	=
48	trials per participant	×
18	participants per glyph and fill type	=
864	**trials in total**	

Fill type was a between subjects factor.

5.2 Hypothesis

Given the results from Experiment 1, and our conjecture on filling, we formulated the following hypothesis.

H1: For the D variation, participants will choose data targets more often than rotated and scaled targets.

H2: Participants will choose data targets for the D variation of the glyph more often than they will for the other variations, irrespective of fill type.

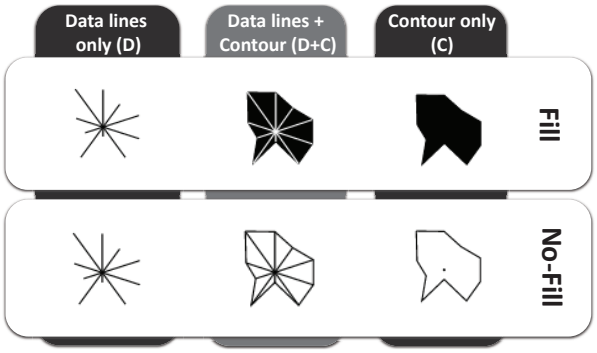

Fig. 7. **Experiment 2 design space**: We enriched the design space from our previous study by adding a "fill" version of the star glyph. The design variations of the first study (i.e., D, $D+C$, C) are applied to both Fill and No-Fill.

H3: Participants will choose the scaled and rotated targets more often than the data targets for the C and $D+C$ variations.
H4: For the filled $D+C$ and C variations, data targets will be chosen less often than for the no-fill variations.
H5: In low dimensional conditions, data targets will be selected more often than other targets irrespective of glyph variation.

5.3 Results

We only report statistically significant results ($p < .05$) for the collected quantitative data. We used a non-parametric Friedman's test for the analysis of the selections between the glyph variations (within-subjects) and a Kruskal-Wallis test for comparisons between glyph designs (between group factor). We did not log completion time, as we could not reliably control pauses during our online experiments.

There was a significant effect of *target types* on the selections made ($\chi^2(2, N = 864) = 149, p < .001$). Overall, participants selected the data target type significantly more often (44.6%) compared to rotated targets (37.3%, $p < .01$), and scaled targets (17.8%, $p < .001$).

For the D variation, included in both experiment groups (fill or no-fill), data targets were selected more often (61.8%) compared to rotated targets (26.4%, $p < .001$) and scaled targets (11.8%, $p < .001$).

For the fill designs (without the D variation), rotated targets were most commonly selected (38.3%), followed by data (35.5%) and scaled ones (25.8%) that were significantly less selected overall (all $p < .05$)

A similar effect is seen for the no-fill variations (without D). Again, rotated targets were most commonly selected (47.2%), followed by data (36.5%) and scaled (16%) ones, with scaled once again being significantly less selected than the other two (all $p < .001$).

In our further analysis we treat each target type as a separate dependent variable (Fig. 8).

Star glyph variations: There was a significant effect of *contour variation* on *data target type* ($\chi^2(2, N = 288) = 32, p < .001$), on *rotated target type* ($\chi^2(2, N = 288) = 12.8, p < .01$), and on *scaled target type* ($\chi^2(2, N = 288) = 7.6, p < .05$).

Post-hoc tests revealed significantly higher selection rates for data targets in variation D (61.8%) compared to $D+C$ (36.5%, $p < .001$) and C (35.4%, $p < .001$) for both fills. Rotated targets were selected significantly less in variation D (26.4%) compared to $D+C$ (44.8%, $p < .001$) and C (40.6%, $p < .05$), while scaled ones significantly less in variation D (11.8%) compared to C (23.3%, $p < .01$).

There was also an effect of *dimensionality* on *data target type* ($\chi^2(1, N = 432) = 32, p < .001$), on *rotated* ($\chi^2(1, N = 432) = 26.1, p < .001$), and on *scaled target* ($\chi^2(1, N = 432) = 8.3, p < .01$).

Participants working with low dimensionalities selected the data target type significantly more often (64.1%) compared to the high dimensional condition (25%, $p < .001$) across all designs. In the high dimensional condition participants selected the rotated (48.4%) and scaled (26.6%) target type significantly more often compared to the low dimensional condition (26.2%, $p < .001$ and 9%, $p < .01$). More details on dimensionality are reported for each fill type later on.

Fill vs. No-Fill Star Glyphs: We consider variation D neither as fill nor as no-fill (common across both experiment groups) and remove it from the analysis. Comparing the fill and no-fill variations we found a significant effect of *filling types* on *rotated* ($\chi^2(1, N = 144) = 4.8, p < .05$), and on *scaled target type* ($\chi^2(1, N = 144) = 8.2, p < .01$).

Post-hoc tests revealed a significantly higher selection rate for the scaled target type for fill designs (25.7%) compared to no-fill (16%, $p < .001$) and for the rotated target type for no-fill designs (47.2%) compared to fill (38.2%, $p < .05$).

No-Fill Star glyphs: The No-Fill star glyphs showed a significant effect of *contour variation* on *data target type* for both low ($\chi^2(2, N = 72) = 8.21, p < .05$) and high dimensional cases ($\chi^2(2, N = 72) = 28.25, p < .001$).

Post-hoc tests revealed a significantly higher selection rate for data target type in variation D for the low and high dimensional case (75%; 62.5%) compared to $D+C$ (61.1%; 15.3%, all $p < .05$) and C (59.7%; 9.7%, all $p < .01$).

The No-Fill star glyphs also showed a significant effect of *contour variation* on *rotated target type* for both low ($\chi^2(2, N = 72) = 7.7, p < .05$) and high dimensional cases ($\chi^2(2, N = 72) = 14.6, p < .001$).

Post-hoc tests revealed a significantly higher selection rate for rotated target types for both the low and high dimensional case in variation C (30.6%, 59.7%) and $D+C$ (29.2%, 69.4%) compared to D (16.7%, 27.8%) (all $p < .05$).

Filled Star glyphs: The filled star glyph had a significant effect of *contour variation* on *data target type* in the high dimensional case ($\chi^2(2, N = 72) = 17.33, p < .001$), and on *scaled target type* in the high dimensional case ($\chi^2(2, N = 72) = 8.5, p < .05$).

Participants working with variation D in high dimensions selected the data target type significantly more often (41.7%) compared to $D+C$ (11.1%, $p < .001$) and C (9.7%, $p < .001$).

The scaled target type was selected significantly more often with variation $D+C$ (43%) and C (40.3%) compared to D (20.8%, $p < .01$ and $p < .05$) in high dimensions.

Variation D: We looked at variation D which is common across fill and no-fill conditions, and found that data targets were selected significantly more in the no-fill (62.5%) than the fill condition (41.6%, $p < .5$). Further analysis shows this is likely due to the order of presentation: in the fill condition, when D was the first design seen, data targets were selected more often (50%), than when they followed another fill design (35%). We explain this in our discussion.

5.4 Discussion

Independent of the fill type, participants using the D glyph variation selected the data target as more similar significantly more often than any other type, giving strong evidence that glyphs without contours promote data similarity comparison rather than shape (H1). Moreover, variation D was the one that the data target was most commonly selected compared to contour variations $C, D+C$ irrespective of fill type (H2).

On the other hand, the most selected targets in contour variations $C+D, C$ were indeed either rotated or scaled variations of the stimulus (H3). This reinforces our findings from the first study, that factors enforcing perceptual unity of shape [18], such as contour containment lead viewers to naturally make shape judgments of similarity rather than data, while open variations of the glyphs lead to similarity choices closer to data comparisons, even without being told what similar means. Also, although not statistically significant, the $C+D$ variation tended to have on average more data target selections than simple C.

The above effects are due mainly to the high dimensional condition. In the low dimensional condition, across all glyph designs, data targets were the ones more often select than all other target types (H5).

When comparing filling types we could not prove that filled star glyphs promote shape judgments more strongly than no-fill star glyphs. Nevertheless, in the fill condition, when the common data-lines design D appeared after fill designs, data selections dropped. We hypothesize

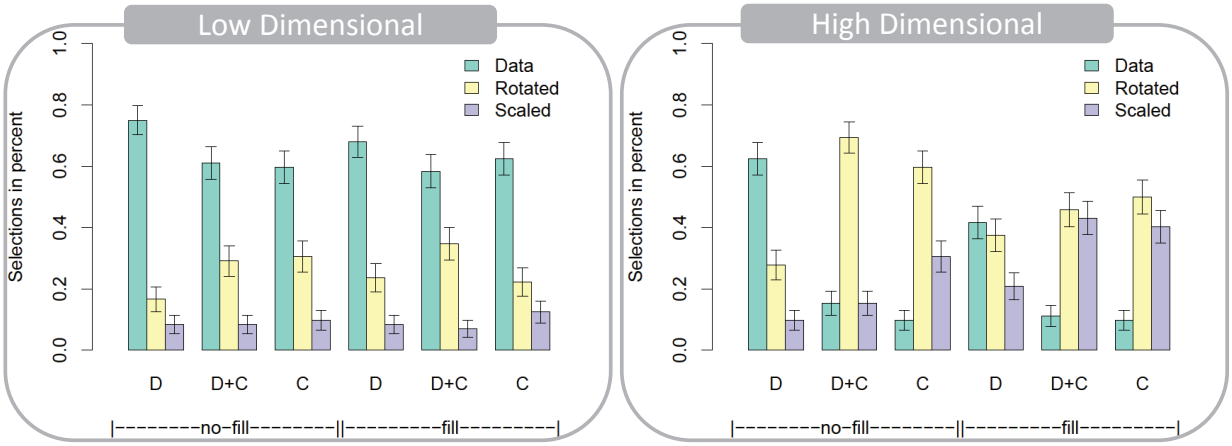

Fig. 8. **Experiment 2 results**: The bar charts illustrate the percentage of selections and the standard deviation for each factor. The left chart represents low dimensionality, the right one the high dimension condition. Even without training or explaining the visual encoding participants using variation D judged data similarity rather than shape similarity.

that seeing a fill design first put participants in a frame of mind of making shape rather than data judgments, a behavior they carry on to the D design that otherwise promotes data similarity. Nevertheless, we saw no significant difference for the variations $C+D, D$ that can actually hold fill color.

Thus, contrary to hypothesis H4, there was no difference in the selection of data targets across fill type. In our experiment the stronger figure and ground distinction, that in the past has been shown to promote unity of shape [18] did not have a noticeable effect in data selections. Perhaps, this finding is also related to the fact that the brain relates surface fill color largely to edge contrast information [34]. Yet, the nature of this perceptual phenomenon does warrant further research in general as the fill type did affect which shape-related similarities people chose. Rotated target types were selected more often with no-fill star glyphs, whereas participants using fill star glyphs more frequently selected scaled target types.

We note again that in this study participants were never told that they were viewing data visualizations, they were just asked to find the most similar glyphs without further instructions. Thus, our results indicate the natural tendency of people to judge glyphs instinctively in a more "data-centric" manner in low dimensionalities, and in high ones when factors that enforce coherent shapes are absent. It is clear that with training we can further enforce data similarity judgments–but given that some glyphs and glyph variations seem to be naturally well suited for data judgments, we focus on those designs and try to further improve their performance with small design variations.

6 Experiment 3: Improvements for Star Glyph

The first experiment showed that people judge data similarity with non-contour designs more accurately while the second experiment showed that non-contour designs also lead to data similarity judgments to be made more naturally. Yet, accuracy in the high-dimensional case was quite low for all main design variations we tested previously. In this last experiment, we thus explore whether we can improve the accuracy of data similarity judgments by adding simple reference structures—tickmarks and grids—to the designs. We focused on static reference structures to learn how much these general approaches would aid data comparison before considering the design of interactive aids.

6.1 Star Glyph Reference Structures

Reference structures such as grids and tickmarks are frequently recommended for data charts to aid in relating content to axes [19]. We, thus, hypothesized that they could provide similar reading aids for star glyphs despite their smaller footprint. Tickmarks and grids use two different types of reference mechanisms. While tickmarks add information to each individual data line only, grids connect the overall glyph design.

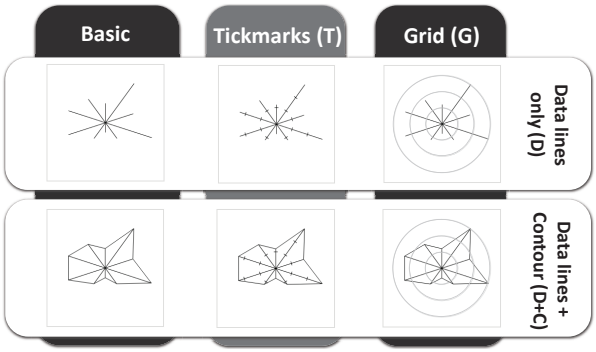

Fig. 9. **Experiment 3 design space**: We have chosen the star glyph only with data whiskers (D) and with an additional contour line ($D+C$) and applied tickmarks (T) and gridlines (G) to these designs.

While there are many different ways to draw grids and tickmarks we settled on the following designs:

Tickmarks T: Whenever a data line exceeds a certain threshold we draw a short orthogonally oriented tickmark on the data lines using the same stroke color. Tickmarks are spaced to be 17 pixels apart. The resulting $D+T$ glyph (see Fig. 9) resembles the snowflake glyph previously mentioned in literature [3] and is also close to how tickmarks are used on axes in many data charts.

Grid G: We draw three circles in the background of the glyph using a gray value of #ccc in RGB color space chosen according to design considerations by Bartram et al. [2]. The circles are spaced 16.6 pixels apart. The resulting design resembles radar graphs or spider plots [31]. As an alternative we considered drawing a gridline at the end of each data line. Doing so would create an underlying texture that could help to identify the overall data distribution across all dimensions. Yet, we chose not to use this design as this texture can be misleading since rotated star glyphs with similar data values would have the same texture, although they have entirely different data values.

Of course, the readability of glyphs could further be improved by adding double encodings (e.g., additionally using color to distinguish dimensions or data values), dimension ordering [25], or sorting the glyphs on the display. Yet, all of these encodings have limitations: use of color is limited to glyphs with a small number of dimensions, dimension ordering may not improve legibility for a large number of variable glyphs in a small-multiple setting, and sorting glyphs may disrupt a pre-defined layout based on other meta-data such as time. We, thus, did not consider these encodings for the study.

6.2 Design and Procedure

Glyphs: We tested the two star glyph variations that performed best in the first experiments: the data-only glyph (D) and the star glyph with data lines and a contour line ($D+C$). The reason for discarding the contour only design (C) is the bad performance for previous similarity judgments, the lack of ability to place tickmarks, and the minimal number of real-world examples of this glyph type in use.

For baseline comparisons we kept the originally tested versions of the star glyph (D, $D+C$) and added two types of reference structures (T, G). The experiment, thus, compared the six different designs (D, $D+T$, $D+G$, $D+C$, $D+C+T$, $D+C+G$) in Fig. 9.

Participants: We recruited 12 data visualization experts (3 female). The age ranged from 23–40 years in age (mean (29.75) & median age (30)). All participants reported normal or corrected-to-normal vision. All experts focused during their studies on data visualization (4 Bachelor; 5 Master; 3 PhD) or a related topic and were familiar with reading data glyphs. They had not participated in the first study.

Task and Procedure: Participants completed data similarity search trials with all 6 designs. The order of the designs was randomized using a latin square. For each design there was a short introduction of the visual encoding and the similarity search task with 5 test questions. The participants had to complete those simple test trials with 80% accuracy in order to continue the experiment. The purpose of the test was to first check the participants' ability to read the visual encoding of the glyph and second to test their data similarity judgments. All participants passed the test section. The introduction was followed by 4 training trials to help the participants develop a strategy for solving the task. For training trials, the correct answer was shown to participants after they had made a choice. Finally the four study trials were shown without any visual feedback of the correct answer.

The experiment took place in a lab setting using a 24" screen with a resolution of 1920×1200 pixels. The experimenter was present during the study. After the study, 11 of the 12 participants filled out a questionnaire for subjective feedback on aesthetics of the designs and strategies used to answer the questions.

Data, Distracters and Dimensionality: Since participants were already $\approx 80\%$ correct in the low dimensional condition in Experiment 1, we only used high-dimensional glyphs in Experiment 3. We generated the data the same way as in Experiment 2 and balanced selection likelihood between distracters. To reduce the chance of a successful random guess we generated only one data point closest in data space (target) and another one second closest in data space (alternative) as in Experiment 1. The experiment included 2 rotated, 2 scaled, 2 random, 1 alternative and 1 target glyph. The stimulus was highlighted and positioned in the middle of the 3×3 matrix as in the two previous experiments. The distracters were randomly arranged around the stimulus.

Overall our experiment was a within-subjects design with the following factors, participants, and trials:

1	glyph (*Star*)	×
2	contour variations (D, D+C)	×
3	improvements (*Basic, T, G*)	×
4	repetitions	=
24	trials per participant	×
12	participants per glyph	=
288	**trials in total**	

6.3 Hypotheses

Based on our previous experiments and the frequent use of reference structures to aid chart reading, we tested the following hypotheses:

H1: Tickmarks (T) in star glyphs improve the accuracy of data similarity judgments for both (D) and ($D+C$) variations compared to the variations without the tickmarks. The additional anchor points help to better read and compare line distances.

H2: An underlying grid (G) in the background of the star glyph provides additional orientation and facilitates more accurate comparison of data values for both (D) and ($D+C$) variations than the variations without the grid.

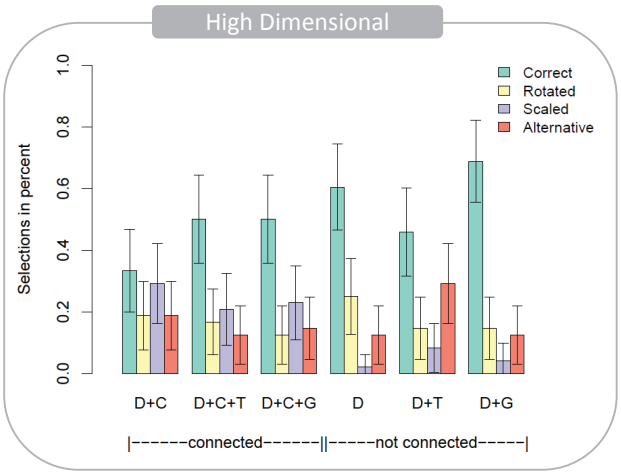

Fig. 10. **Experiment 3 results** of the percentage of selections and the standard deviation for each factor. Design improvements (T, G) do not significantly increase the accuracy of the two star glyph variations ($D+C$, C).

H3: The contour variation $D+C$ benefits more from the additional reference structures than the D variation since contour has previously shown to lead to shape comparison rather than data similarity comparisons.

H4: Completion time is higher for designs enriched with reading marks (T or G). The viewer has to invest more mental effort to process the additional visual information.

6.4 Results

Similarly to Experiment 1 we used a non-parametric Friedman's Test on the data to analyze accuracy, and a one-way ANOVA for the completion time. We only report statistically significant results ($p < .05$).

The overall accuracy was 51.4%, with designs with grids (G) being more accurate (59.4%), followed by the tickmark designs (T) (47.9%) and then designs without additional marks (46.9%). There was a statistical trend for different types of reference structures on accuracy ($p < .1$), with glyphs with grids being more accurate than with tickmarks. There was no difference between designs with reference structures and the baseline design.

Next, we compared the different glyph variations without contour (D) and with contour ($D+C$). As in Experiment 1, participants were significantly more accurate with variation D (60.4%) than when the contour was present $D+C$ (33.3%, $p < .01$).

Reference structures on glyphs without contours (the D glyphs) did not significantly improve accuracy over the glyph without the reference structure. Participants were 60.4% accurate with D, 68.8% accurate with ($D+G$), and 45.8% accurate with ($D+T$). Nevertheless, we note that the mean accuracy of the ($D+G$) variation is indeed higher than for D only. We also found that for the two variations using reference structures, grids ($D+G$) were significantly more accurate than tickmarks ($D+T$) (45.8%, $p < .05$).

For the contour variations, we have a statistical trend ($p < .1$) indicating that the accuracy of both the contour variation with a grid ($D+C+G$) and the one with tickmarks ($D+C+T$) tend to be more accurate (both 50%) than that of simple glyph with contour ($D+C$) with accuracy 33.3% ($p = .06$ and $p = .08$ respectively).

Looking at differences across variations, we also found that $D+G$ (68.8%), which had the highest overall mean accuracy, performed significantly better than $D+C$ (33.3%, $p < .001$) and had a statistical trend to perform better than $D+C+G$ ($p = .1$) and $D+C+T$ ($p = .8$).

The mean number of selections per distracter type are shown in Fig. 10. We found a significant effect of *variation* on *distracter* ($\chi^2(5, N = 48) = 12.68, p < .05$). Participants using variations with contour lines most often selected the scaled distracter (24%) followed by the rotated (16%) and the alternative (15%) distracter. For the non-contour variations participants chose the alternative and the rotated distracter equally often (18%) followed by the scaled distracter (5%).

No significant results can be reported for the completion time, thus we cannot confirm that additional marks influenced comparison times. However, participants needed approx. 2*sec* longer when working with designs using additional marks. Average completion time was 22*sec* per trial ($D = 21.7sec$, $D+G = 24.8sec$, $D+T = 26.1sec$, $D+C = 17.9sec$, $D+C+G = 21.5sec$, $D+C+T = 22sec$).

The questionnaire showed that the glyph variations with contours ranked highly amongst participants' aesthetic preferences. The mostly strongly preferred glyph variation was $D+C+G$ (5/11 participants), followed by $D+C$ (3/11 participants). Interestingly, no participants preferred the D variation even though its mean accuracy (60.4%) was higher than $D+C+G$ (50%). Participants also ranked the D variation as hard to use (median=6 on a 7-point Likert scale) with all other designs ranking at least between median 4–2. The $D+C+T$ and $D+C+G$ variations were both found easy to use (median=2). We report on the results of the questions regarding strategy in our discussion section.

6.5 Discussion

Adding reference structures to the star glyph did not have the effect on accuracy we were expecting for our data similarity search task. Additional anchor points on the data line (i. e., tickmarks) did not significantly improve the comparison of data points. Therefore, we cannot accept H1. Nevertheless, there was a statistical trend indicating that an overall reference in the background (i. e., gridlines) may increase accuracy, especially in the case of contour star glyphs, providing some evidence for H2.

This lack of strong significant effects is surprising, especially given that most participants mentioned in the questionnaire that for the simple star glyph D, gridlines (81%), and to a lesser extent tickmarks (72%), helped them find the most similar data point. Although the mean accuracy for the $D+G$ variation was indeed higher, the effect was not significant, perhaps due to the already very good performance of the D variation. The value of gridlines and tickmarks in general may warrant further research. As Few notes [10], gridlines may be useful only in specific cases, e. g., when small differences have to be compared. Therefore, it is possible that for other tasks, such as direct lookup, these additional reference marks could help more strongly.

For the star glyph with contour ($D+C$), only 54% of our participants reported using tickmarks and 36% gridlines to complete the task. From their reports they felt (erroneously) that glyphs with contours are easier to compare and, thus, did not make conscious use of the additional improvements. Thus, in the contour case, participants were not only more error prone, but also misled to feel confident in their choices, ignoring the marks that could help them improve their performance. Nevertheless, it is highly likely that the addition of reading marks was taken into account, even if unintentionally, explaining the trend we see for both the tickmark and grid variation to be more accurate than simple contour glyphs (H3).

Finally, we could not confirm H4 due to a lack of significant results when comparing task performance time.

Even though participants using variation (D) performed very well, it is interesting that they did not like this design variation. On a 7-step Likert scale 63% of the participants rated the design with either 6 (difficult to use) or 7 (very difficult to use). Most participants (46%) preferred the star glyph with contour and gridlines, with only 1 participant rating it with a 5 (slightly difficult to use) and the others with 3 or better.

Given the results of this experiment the benefit of using reference structures for star glyphs is limited. Especially since in real world scenarios when multi-dimensional glyphs are projected to two dimensional surfaces, there is the possibility of over-plotting, and adding marks or gridlines could worsen this effect due to the additional ink introduced.

7 Design Considerations

With the results gained from the analysis and discussions we derive the following design considerations.

When judging data similarity avoid contours in glyph designs. Viewers have a natural tendency to judge data similarity in star glyphs without contours. In all our experiments viewers were tricked into making shape-based, rather than data-based judgments when using contours. This is especially true if glyphs in the visualization are scaled or rotated versions of each other.

For low number of dimensions (around 4) any glyph variation can safely be used for data similarity judgments. In the first and second experiment viewers naturally leaned towards data similarity for each glyph variation in low dimensions, even without training.

When there is a need for contours, add data lines to the design to strengthen data similarity judgments. Participants independent of glyph design (fill or no-fill) judged data similarity better using the $D+C$ variation compared to C in the first two experiments. Although, there was no statistical significance, mean data comparisons for contour + data variations were always higher than contour only.

When there is a need for contours, the designer can decide whether or not to use fill color. Our Experiment 2 gave no indication that fill color degrades the performance of glyphs with contour.

When clutter is an issue avoid reference structures in non-contour star glyphs for similarity search tasks. Results of Experiment 3 illustrate that even though participants preferred using tickmarks or grids they did not perform significantly better with them, especially for glyphs without contours. Nevertheless, there is a statistical trend that shows that tickmarks and grids improve glyphs with contours.

If references are required use grids rather than tickmarks. Independent from the design (i.e., with or without contour) gridlines always increased mean accuracy, which is not true for tickmarks.

8 Conclusion

We investigated the effect of contours on the perception of similarity for star glyphs. In a first controlled experiment with 24 participants, we examined the influence of contours for novice and expert users. We found that experts can be tricked into making similarity judgments based on shape, rather than data closeness, when viewing glyphs with contours. For novices the effect was less pronounced. To better understand how people naturally judge similarity, we conducted a second online study with 36 participants and asked about intuitive notions of similarity. We found that removing contours and fillings from star glyphs, naturally increased the perception of data similarity (rather than shape) even when viewers were not trained or aware they are viewing data.

As a next step we tried to improve the accuracy in judging data similarity. We added two types of reference structures to the star glyph, gridlines and tickmarks, and tested these alternatives in a third experiment. Surprisingly, the star glyph without contour line and reference structure still performs best for similarity search tasks. Based on our findings we provide a set of glyph design considerations, the most important being that visualization designers should avoid contours when representing similar data points to analysts.

In summary, our work has provided insights as to the effect of contours on similarity perception for star glyphs. Similarity perception is an important task especially since glyphs are mostly used for quick overviews, and to detect trends and similarities [34], rather than to provide highly accurate value representations [13]. Other tasks performed on glyphs, however, such as exact data value reading, may yield different results from ours, e. g., adding grids or tickmarks could improve performance [27].

Given our experimental results and our provided guidelines, we would like to focus on two future research directions. First, we would like to examine whether our findings can be applied to different glyph designs (e. g., profile glyphs [7]), as it is unclear if contours promote shape similarity rather than data in glyphs that already resemble familiar data charts. Thus we can derive a more generalized set of design considerations. Second, based on our results on possible pitfalls in data similarity judgments, we plan to introduce a more task specific training, focusing on rotated and scaled distracters that seem to mislead viewers the most. Given our results, both novices and experts would profit from such specific training, especially when using glyphs with contours.

Acknowledgments

This work has been supported by the Consensus project and has been partly funded by the European Commission's 7th Framework Programme through theme ICT-2013.5.4 ICT for Governance and Policy Modelling under contract no.611688.

REFERENCES

[1] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data*. Springer Verlag, Berlin, 2006.

[2] L. Bartram, B. Cheung, and M. C. Stone. The Effect of Colour and Transparency on the Perception of Overlaid Grids. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1942–1948, 2011.

[3] I. Borg and T. Staufenbiel. Performance of Snow Flakes, Suns, and Factorial Suns in the Graphical Representation of Multivariate Data. *Multivariate Behavioral Research*, 27(1):43–55, 1992.

[4] R. Borgo, J. Kehrer, D. H. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications. In *Proceedings of Eurographics*, pages 39–63. Eurographics, 2012.

[5] N. Boukhelifa, A. Bezerianos, T. Isenberg, and J.-D. Fekete. Evaluating Sketchiness as a Visual Variable for the Depiction of Qualitative Uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2769–2778, Dec. 2012.

[6] Y.-H. Chan, C. Correa, and K.-L. Ma. The Generalized Sensitivity Scatterplot. *IEEE Transactions on Visualization and Computer Graphics*, 19(10):1768–1781, Oct. 2013.

[7] S. H. Du Toit, A. G. W. Steyn, and R. H. Stumpf. *Graphical Exploratory Data Analysis*. Springer-Verlag, New York, 1986.

[8] J. Elder and S. Zucker. The Effect of Contour Closure on the Rapid Discrimination of Two-Dimensional Shapes. *Vision Research*, 33(7):981–991, 1993.

[9] J. H. Elder and S. W. Zucker. Evidence for Boundary-Specific Grouping. *Vision Research*, 38(1):143–152, 1998.

[10] S. Few. Grid Lines in Graphs are Rarely Useful. In *DM Review*. Perceptual Edge, Feb. 2005. `http://www.perceptualedge.com/articles/dmreview/grid_lines.pdf`.

[11] F. Fischer, J. Fuchs, and F. Mansmann. ClockMap: Enhancing Circular Treemaps with Temporal Glyphs for Time-Series Data. In *Short Paper Proceedings of the Eurographics Conference on Visualization (EuroVis)*, pages 97–101. Eurographics, 2012.

[12] M. Friendly. A.-M. Guerry's "Moral Statistics of France": Challenges for Multivariable Spatial Analysis. *Statistical Science*, pages 368–399, 2007.

[13] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of Alternative Glyph Designs for Time Series Data in a Small Multiple Setting. In *Proceedings Human Factors in Computing Systems (CHI)*, pages 3237–3246. ACM, 2013.

[14] P. Garrigan. The Effect of Contour Closure on Shape Recognition. *Perception*, 41(2):221–235, 2012.

[15] J. Heer and M. Bostock. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings Human Factors in Computing Systems (CHI)*, pages 203–212. ACM, 2010.

[16] C. Kintzel, J. Fuchs, and F. Mansmann. Monitoring Large IP Spaces with Clockview. In *Proceedings Symposium on Visualization for Cyber Security*, page 2. ACM, 2011.

[17] A. Klippel, F. Hardisty, and C. Weaver. Star Plots: How Shape Characteristics Influence Classification Tasks. *Cartography and Geographic Information Science*, 36(2):149–163, 2009.

[18] K. Koffka. *Principles of Gestalt Pschology*. Lund Humphries, London, 1935.

[19] S. Kosslyn. *Graph Design for the Eye and Mind*. Oxford University Press, New York, 2007.

[20] M. Lee, R. Reilly, and M. Butavicius. An Empirical Evaluation of Chernoff Faces, Star Glyphs, and Spatial Visualizations for Binary Data. In *Proceedings of the Asia-Pacific Symposium on Information Visualisation (APVis)*, pages 1–10. Australian Computer Society, Inc., 2003.

[21] P. A. Legg, D. H. Chung, M. L. Parry, M. W. Jones, R. Long, I. W. Griffiths, and M. Chen. MatchPad: Interactive Glyph-Based Visualization for Real-Time Sports Performance Analysis. In *Computer Graphics Forum*, volume 31, pages 1255–1264, 2012.

[22] J. Meyer-Spradow, L. Stegger, C. Doring, T. Ropinski, and K. Hinrichs. Glyph-based SPECT Visualization for the Diagnosis of Coronary Artery Disease. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1499–1506, Nov./Dec. 2008.

[23] OECD Better Life Index. Your better life index, 2011. `http://www.oecdbetterlifeindex.org/`. Retrieved July 2013.

[24] S. E. Palmer. *Vision Science: Photons to Phenomenology*. The MIT Press, London, 1999.

[25] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter Reduction in Multi-Dimensional Data Visualization Using Dimension Reordering. In *Symposium on Information Visualization (InfoVis)*, pages 89–96. IEEE, 2004.

[26] R. M. Pickett and G. G. Grinstein. Iconographic Displays for Visualizing Multidimensional Data. In *Proceedings of the Conference on Systems, Man, and Cybernetics*, volume 514, page 519. IEEE, 1988.

[27] N. B. Robbins. *Creating More Effective Graphs*. Wiley-Interscience, Hoboken, NJ, USA, 2005.

[28] J. Saarinen and D. M. Levi. The Effect of Contour Closure on Shape Perception. *Spatial Vision*, 12(2):227–238, 1999.

[29] J. Siegel, E. Farrell, R. Goldwyn, and H. Friedman. The Surgical Implications of Physiologic Patterns in Myocardial Infarction Shock. *Surgery*, 72(1):126, 1972.

[30] M. Stefaner. Wahlland, 2013. `http://moritz.stefaner.eu/projects/wahlland/`. Retrieved March 2014.

[31] G. Von Mayr. *Die Gesetzmässigkeit im Gesellschaftsleben*, volume 23. R. Oldenbourg, Munich, 1877.

[32] M. Ward. Multivariate Data Glyphs: Principles and Practice. *Handbook of Data Visualization*, pages 179–198, 2008.

[33] M. O. Ward. A Taxonomy for Glyph Placement Strategies for Multidimensional Data Visualization. *Information Visualization*, 1(6):194–210, 2002.

[34] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, Waltham, 2012.

[35] L. Wilkinson. An Experimental Evaluation of Multivariate Graphical Point Representations. In *Proceedings Human Factors in Computing Systems (CHI)*, pages 202–209. ACM, 1982.