Glanceable Visualization: Studies of Data Comparison Performance on Smartwatches

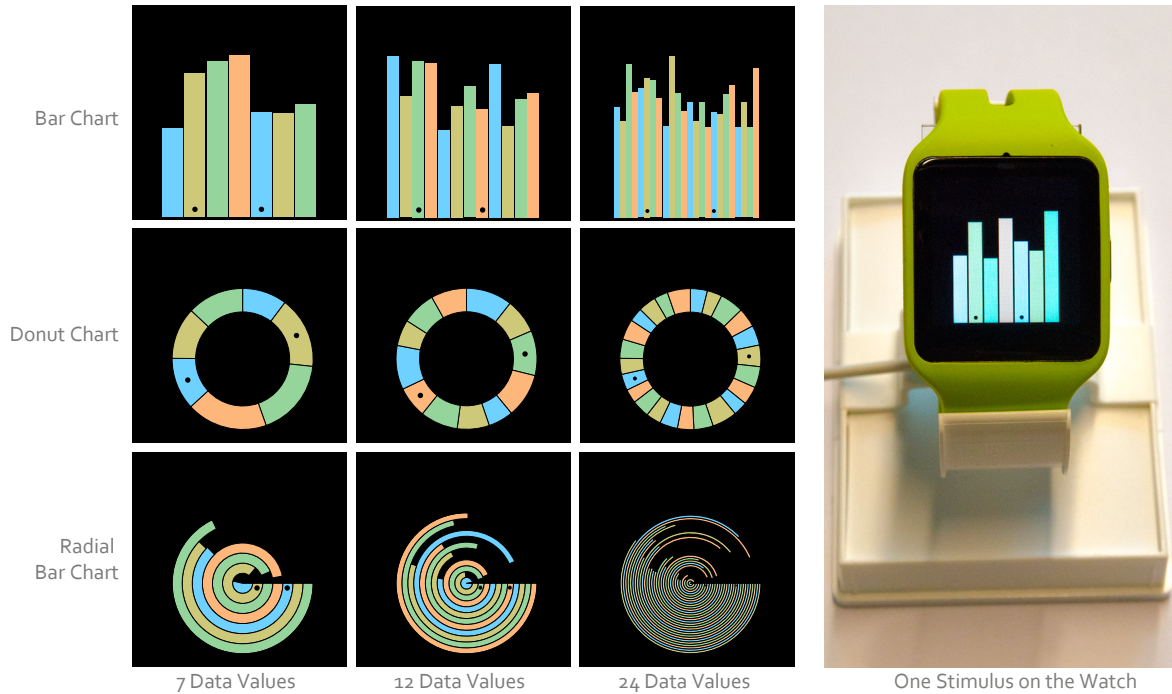Tanja Blascheck, Lonni Besançon, Anastasia Bezerianos, Bongshin Lee, and Petra Isenberg

Fig. 1. Left: Example images of the stimuli used in the two perception studies. We tested three *chart types* (Bar, Donut, and Radial) with three *data sizes* (7, 12, and 24) on a smartwatch. When printed without rescaling the image, the sizes correspond to the size of the stimuli as shown on our smartwatch (28.73 mm × 28.73 mm). Right: One stimulus shown on the smartwatch.

Abstract—We present the results of two perception studies to assess how quickly people can perform a simple data comparison task for small-scale visualizations on a smartwatch. The main goal of these studies is to extend our understanding of design constraints for smartwatch visualizations. Previous work has shown that a vast majority of smartwatch interactions last under 5 s. It is still unknown what people can actually perceive from visualizations during such short glances, in particular with such a limited display space of smartwatches. To shed light on this question, we conducted two perception studies that assessed the lower bounds of task time for a simple data comparison task. We tested three chart types common on smartwatches: bar charts, donut charts, and radial bar charts with three different data sizes: 7, 12, and 24 data values. In our first study, we controlled the differences of the two target bars to be compared, while the second study varied the difference randomly. For both studies, we found that participants performed the task on average in <300 ms for the bar chart, <220 ms for the donut chart, and in <1780 ms for the radial bar chart. Thresholds in the second study per chart type were on average 1.14–1.35× higher than in the first study. Our results show that bar and donut charts should be preferred on smartwatch displays when quick data comparisons are necessary.

Index Terms—Glanceable visualization, smartwatch, perception, quantitative evaluation, data comparison.

✦

1 INTRODUCTION

- *Tanja Blascheck and Petra Isenberg are with Inria. E-mail: {tanja.blascheck, petra.isenberg}@inria.fr*
- *Lonni Besançon is with Université Paris Saclay. E-mail: lonni.besancon@gmail.com*
- *Anastasia Bezerianos is with Université Paris Sud, Inria, CNRS and Université Paris Saclay. E-mail: anastasia.bezerianos@lri.fr*
- *Bongshin Lee is with Microsoft Research. E-mail: bongshin@microsoft.com*

Smartwatch use in the US is expected to rise by 60% in 2019, leading to ~15% of consumers owning a smartwatch [45]. Smartwatches are in particular used to monitor and track activities or respond to notifications [32, 37, 45, 48]. Recent field studies of smartwatch usage show that a majority of smartwatch interactions involve quick glances or peeks at the smartwatch [37], which were often shorter than 5 s. These quick glances limit the amount of information a viewer can take in. This suggests that smartwatch interfaces need to be suitable for these quick peeks and be designed to convey information as quickly as possible.

Visualizations, by design, allow to effectively convey information and reduce reading time [8]. Therefore, they may be able to convey a large amount of information during brief glances at a smartwatch. Many possible application contexts exist, in which visualizations can help to provide detailed information during quick glances. For example,

in the context of personal data tracking, visualizations on smartwatches can show workout progress, can be a fitness and health indicator for runners, or show elevation profiles during a hike in the mountains.

While visualizations for smartwatches have potential benefits, the small display size as well as people's viewing behavior (i. e., quick glances) pose unique design challenges. Recently, research has studied small-scale visualizations, for example, word-sized graphics [2, 19] and data glyphs [14]. However, this body of work does not consider how quickly viewers can read small-scale visualizations. There are no guidelines in visualization for designing small-scale visualizations that consider both small display size as well as brief viewing times.

We present the design and results of two perceptual studies conducted to assess how quickly viewers can read information from small-scale visualizations shown on a smartwatch. To conduct these studies, we first empirically derived metrics on how people position and orient a smartwatch when reading information. Next, we designed a two-alternative forced choice experiment [30], in which we fixed the smartwatch at the derived average viewing angle and distances (cf. Fig. 1, right). Participants performed a simple data comparison task, during which visualizations were shown using a staircase procedure that varied the stimulus exposure duration [30]. We tested this task for three *data sizes*: *7*, *12*, and *24* data values; and three basic *chart types* commonly used on smartwatches: Bar 📊, Donut ⭕, and Radial 🌈 (cf. Fig. 1, left). In a first study, the compared targets had a controlled size difference of 25%, while we used randomized data for the size of the targets in the second study. For both studies, we calculated a reading time threshold for each *chart type* × *data size* combination to target ~91%-correct responses. For the first study, our results show that bar chart and donut chart thresholds were low across all *data sizes*: 159 ms for the donut and 245 ms for the bar chart on average. The average for the radial bar chart was much larger at 1548 ms. In the second study, we found a time threshold of 285 ms for the bar chart, 216 ms for the donut chart, and 1772 ms for the radial bar chart which were 1.16×, 1.35×, and 1.14× higher respectively than the results from the first study. These thresholds are still low and allow us to conclude that bar and donut charts should be the preferred encodings when quick comparisons are necessary.

The key contributions of this work are threefold. We present the first controlled experiments on graphical perception of three *chart types* on smartwatches, enabled by a study to understand how people position and orient smartwatches to read information. We also make a methodological contribution: we build on previous work in psychophysics and studies of the perception of visual stimuli [30] to offer the first staircased visualization perception study conducted on smartwatches. Finally, we reflect on our results and discuss future research directions.

2 RELATED WORK

Considerable research on smartwatches has been conducted in the field of human-computer interaction (HCI). Much of this research concerned input modalities (touch, gestures, etc.) [13, 33]. We focus on visual perception on smartwatches rather than input; therefore, we do not discuss this research further. Instead, we focus on studies of smartwatch use in the wild and what was found about the length of smartwatch usage sessions. We further discuss related work from the visualization community on small-scale visualizations and their evaluation, as well as previous work on smartwatch visualization. We cover background information relevant to our study design in Sect. 4.

2.1 Smartwatch Use in the Wild

Previous studies in the field of HCI aimed to quantify smartwatch use. In an online questionnaire of 59 smartwatch users, Min et al. [32] found that people considered the following smartwatch features to be the most important: information on current time, notifications, and sports tracking. Pizza et al. [37] recorded smartwatch use of 12 participants using wearable cameras, and collected 1009 instances of smartwatch use on video. In approximately half of all instances, participants used the smartwatch to check time for an average duration of ~2 s. Notifications with an average duration of ~7 s were the second most frequently

used smartwatch feature (~17% of all instances). Instances of smartwatch use across all types of applications were ~7 s long. In another study, Visuri et al. [48] logged smartwatch use of 307 participants and analyzed two types of smartwatch interaction: *peeks* (≤5 s) and *interactions* (>5 s). They also differentiated user-initiated sessions (~80% of all sessions) and notification-initiated sessions. The authors recorded an average usage time of ~8 s for user-initiated sessions and ~11 s for notification-initiated sessions. More than half of all sessions were peeks. These results suggest that quick unprompted peeks at the smartwatch were the most common type of smartwatch use. In our work, we offer an alternative step towards establishing if data visualizations can be effectively perceived during such short peeks. It should be noted, however, that the authors in the latter two studies did not differentiate between the time a smartwatch functionality was visible and the time a viewer spent actively attending to the shown content. As such, peeks at a smartwatch could be even shorter.

2.2 Micro Visualizations

Typical smartwatches (as of March 2018) have a resolution between 128–480 px per side with a viewable area of around 30–40 mm. The best selling smartwatch in March 2017 according to Amazon.com was the Fitbit Ionic with 348 px × 250 px on a 1.4 in screen (29.3 mm × 21 mm) at 302 PPI (pixel per inch). Given the small number of pixels, visualizations for smartwatches are necessarily restricted in the amount of content they can show. Visualization research in the past has considered this problem of small-scale visualizations under three topics: micro visualizations, data glyphs, and word-scale visualizations.

Parnow [35] defined micro visualizations as data representations that are small in physical display space, used in the context of text documents, and only encode a few data dimensions. This definition is closely related to sparklines [46] and word-scale visualizations [18], both of which are word-sized graphics used to accompany text. We critique Parnow's usage of the term because it restricts the general term *micro visualization* to the context of text documents and limited data dimensions while terms with similar usage contexts already exist in the literature [18, 46]. In addition, a year earlier, Brandes [6] already used the term to propose an important research direction in visualization: micro visualizations that are high-resolution visualizations for small to medium-sized displays. Brandes argues that micro visualizations allow to display data in eye span and — when properly designed — also allow to read on a micro (detail) and macro (overview) level. Ultimately, visualizations for smartwatches will benefit from these design principles. However, it is still unknown which amount of representation complexity viewers can perceive at a small scale and in particular during quick peeks or glances. Our study, therefore, focuses on simple visualizations that are assessed under a simple data comparison task. This knowledge can serve as a benchmark for determining the perceptible visual complexity of smartwatches.

Researchers have previously proposed specific micro visualizations for small physical display spaces. Horizon Graphs [39] are one such example. This time series visualization achieves a small display space by cutting filled line charts into bands, coloring the bands, and layering them on top of each other. Scented Widgets are an example of simple chart-type micro visualizations embedded in the context of GUI widgets to provide information scent [51]. Small graphics can also serve to enrich and accompany text. Tufte [46] defined sparklines as *small, intense, simple, word-sized graphics with typographic resolution*. Goffin et al. expanded on Tufte's definition and proposed the term word-scale visualization [20] to define a wider variety of small, embedded data-driven graphics. Several different word-scale visualizations have been introduced in the past. Beck et al. [2] recently presented an overview of word-sized graphics for scientific texts citing many examples including SportLines [36], GestaltLines [7], Separation Plots [24], or the use of in-situ visualizations for showing eye tracking results [1]. In the latter work, the authors tried to systematically transfer desktop-sized visualization techniques to word-scale visualizations and found that many large-sized visualizations could be scaled to smaller versions. However, we are still missing guidelines on how to do this systematically.

Data glyphs are another type of data representation, that has been

associated with a small display footprint. Data glyphs encode single data points individually by assigning their data dimensions to one or more marks and their corresponding visual variables. According to Fuchs [14], researchers typically embed glyphs in a meaningful context-giving layout, but they are, by definition, not restricted in display size. In practice, however, glyphs are often displayed at a small scale in a small multiples setting. Glyphs are one of the most relevant types of related work for micro visualizations because researchers often aim at their design being holistically perceivable by a data shape — the macro reading Brandes [6] referred to — and they can be useful even at a small scale. We consider small data glyphs as a specific type of micro visualization that encode multiple data dimensions, are embedded with a meaningful layout, and often include minimal or no reference structures such as grid lines, labels, legends, or specific data axes. In the last 60 years, many glyph designs have been published and many of them have been empirically evaluated [5, 15, 49, 50].

2.3 Studies of Micro Visualizations

We are aware of only few studies that have compared physical display size for small data representations [26, 29, 36]. Heer et al. [26], for example, provide a first indication that studying micro visualizations of different size may lead to unexpected results compared to what we know about larger-sized data representations. In a comparison of filled line charts and Horizon Graphs, the authors found that small chart heights negatively affected accuracy and speed of data comparison and that smaller size had a greater impact on the filled line charts than on the Horizon Graphs. This is surprising because line charts are a familiar technique that has previously been shown to be quickly and accurately readable. Similarly, the aspect ratio selection for specific types of charts [43] or the choice of color to encode categorical variables [42] have been shown to be similarly sensitive to changing display sizes.

Studies on word-scale visualizations are still rare. In their work on SportLines, Perin et al. [36] tested several design alternatives using varying display sizes (ranging from 20×15 to 80×60 px). The authors did not control for the dots per inch (DPI) of participants' screens; therefore, the results can only be compared relative to one another because different DPIs result in different physical display spaces. When comparing representations, participants always ranked the smallest designs the least preferred.

Fuchs et al. recently presented a systematic review of 64 quantitative study papers on data glyphs [15]. In this paper, the authors characterized previous studies on data glyphs according to the tested designs, study questions, as well as data and tasks used during the studies. In addition, the authors synthesized the study results, discussed trade-offs in the data glyph design space, and indicated which research questions remain open. For example, they found no studies on data glyphs that specifically used display size or viewing time as a study factor.

2.4 Visual Design for Smartwatches

Little research so far was specifically dedicated to visualizations on smartwatches. Chen [9] recently presented a visualization system for exploring time-series data. The system uses the border of the screen to show data overviews and the central region for detailed information. Horak et al. [28] studied the combination of smartwatch and large display data exploration. The smartwatch served as a personalized toolbox, which allowed interaction on simple data representations. In these projects, smartwatches are interacted with and looked at for prolonged time periods, unlike the short glances that are common for other types of applications [37] and that we focus on here. Glanceable visual feedback was instead the focus of past work by Gouveia et al. [22]. The authors integrated small representations of activities as part of the smartwatch face so feedback was available as part of the activity of checking the time. The authors observed different types of behavior change due to the glanceable feedback provided. Instead of behavior change, we are interested in how much information viewers can assess "at a glance" and to ultimately provide guidance on the design of glanceable visualizations.
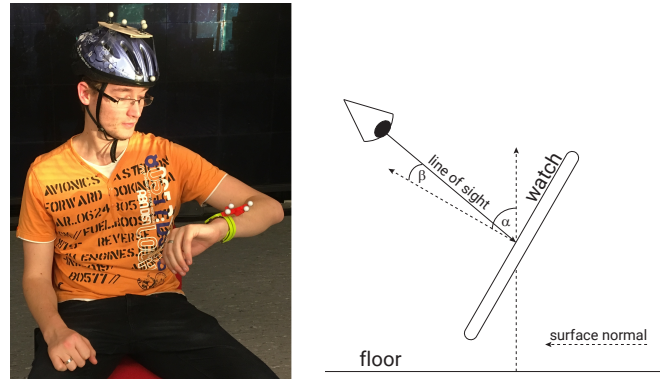

Fig. 2. Left: A participant, wearing helmet and smartwatch both with markers attached, is reading a text displayed on the smartwatch while sitting. Right: Measurements taken during the pre-study.

3 PRE-STUDY: POSITION AND ORIENTATION OF WRISTWORN SMARTWATCHES FOR READING TASKS

In a pre-study, we investigated at which viewing angle and distances smartwatches are commonly held while people are reading information on them. The full details of this study can be found in our previous workshop paper [4]. Here, we summarize the most important findings.

The study used a within-subjects design with one factor: whether participants were sitting or standing. We hypothesized that a seated position would potentially change the viewing angle and distances at which a smartwatch would be held while reading information.

During the study, participants wore the same smartwatch used in the main study, but with attached tracking markers. They also wore a firmly attached bike helmet with tracking markers (cf. Fig. 2, left). Participants read 20 sentences per condition on the watch. While participants were reading, we captured position and orientation for helmet and smartwatch using a 6-camera 3D real-time Vicon tracking system. The Vicon tracking system logged for both the helmet and the smartwatch: a timestamp, an x-, y-, and z-coordinate for the position, and the orientation of the object as a quaternion (qx, qy, qz, qw).

From this data, we calculated: (1) *Pitch angle*, the angle between the watch's normal and the floor's normal (angle α in Fig. 2, right); (2) *Smartwatch distance*, the distance from the center between both eyes to the smartwatch's center, corresponding to the length of the *line of sight* (LOS) in Fig. 2, right; (3) *LOS offset*, the angle between LOS and the inverted smartwatch face's normal (=angle β in Fig. 2, right).

The results showed that average pitch angles were 48° ($SD = 15°$) when sitting, and 52° ($SD = 13°$) when standing. This leads to an average pitch angle of 50° ($SD = 14°$) for both conditions combined. For calculating the LOS offset, we removed 23 (4%) out of 480 trials as outlier trials, i. e., all trials beyond 2 SD per participant. We found that neither sitting nor standing participants' LOS aligned with the smartwatch face's normal, indicating a slightly tilted view. The LOS offset was 11° ($SD = 8°$) when sitting and 9° ($SD = 6°$) when standing. This gives us an average angle of 10° ($SD = 8°$) between the two conditions. Finally, the watch distance was 27.6 cm ($SD = 3$ cm) when sitting, and 28 cm ($SD = 5$ cm) when standing. The average distance between the two conditions is 28 cm ($SD = 5$ cm) .

4 STUDY: DATA COMPARISON TIME ON A SMARTWATCH

With the results from our pre-study, we set up a controlled experiment to find a minimum *time threshold* participants would need to conduct a simple data comparison task on different small-scale visualizations on a smartwatch. We investigated three *chart types* (Bar, Donut, Radial) and three *data sizes* (*7*, *12*, and *24*).

4.1 Study Design

After a review of the top ten Android Wear apps listed in the Google Play Store with a fitness, health, weather, financial and business context, we focused on the three most common *chart types* we encountered: bar
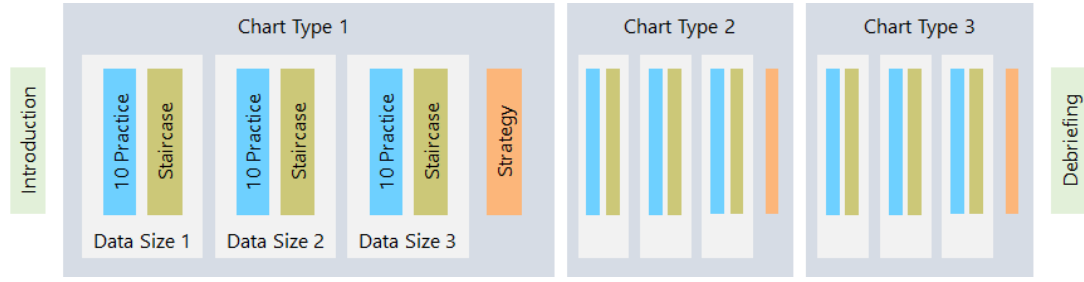
Fig. 3. Procedure of one study session. After the introduction, participants performed nine staircases (for 3 *chart types* × 3 *data sizes*). Between each *chart type* we asked participants about their strategy. The study ended with a debriefing session.

charts, donut charts, and radial bar charts. We tested each chart with three different *data sizes* (*7*, *12*, and *24* data values). We chose these three *data sizes* because most of the apps we found displayed data over time. Common temporal views involve 7 days per week, 12 months per year, or 24 hours per day. We initially also considered 30 data values (e. g., 30 days per month), but with our chosen *chart types* this led to illegible charts on the smartwatch. In summary, the study consisted of 9 conditions: 3 *chart types* × 3 *data sizes* (cf. Fig. 1, left).

In our study, we used a two-alternative forced choice (2AFC) design [23,30], a popular technique in psychophysics to assess the perception of visual stimuli. This technique generally involves showing two alternative options per trial and asking participants to select one of the two options based on a given task. In our case, each stimulus involved a *chart type* × *data size* combination. On each chart, we highlighted two target data values using black dots. Participants' task was to select the larger of the two targets. Four intervening images to account for after images as done by Greene and Oliva [23] followed. We showed each stimulus for a specific *stimulus exposure duration* to be able to study the average task duration participants need to perform this task.

The stimulus exposure duration was adapted w.r.t. participants' responses. The goal of a method that adapts a variable of interest is to find the threshold of the psychometric function [30], a function that describes, in our case, the relationship between stimulus exposure duration and the forced-choice responses of a human observer. The threshold of this function represents the average task duration at which participants can still perform the task with a given percentage of correct responses. This percentage of correct answers depends on the exact type of adaptive approach chosen.

We used a *weighted up/down staircase* method, in which we decreased the stimulus exposure duration by 300 ms after three correct responses and increased the stimulus exposure duration by 100 ms after each error. Before the first error, we decreased the stimulus exposure duration by 300 ms to reach the true threshold for a participant more quickly. This procedure generally leads to a threshold that represents ~91% correct responses [16]. For each condition, participants performed a series of trials until one of two termination criteria was reached: (1) 15 reversals of decrease-increase and increase-decrease or (2) 150 trials in total. Throughout this paper, we refer to a block of trials per condition as one staircase. We counterbalanced the order of *chart types* and the order of *data sizes* per participant, both using a Latin square design.

4.2 Procedure

Participants performed nine staircases (for 3 *chart types* × 3 *data sizes*; cf. Fig. 3). Each session started with participants first signing a consent form and filling out a questionnaire about their background information. Next, participants read a short description of the study, an overview about the different conditions, and the general procedure of one trial on paper. For each staircase, participants performed ten practice trials to familiarize themselves with the general procedure. After finishing the practice trials, participants continued with the staircase until they reached one of the two termination criteria. Fig. 4 shows the general procedure for one trial. Each trial began with participants pressing a button and the smartwatch showing if the answer was correct or not.
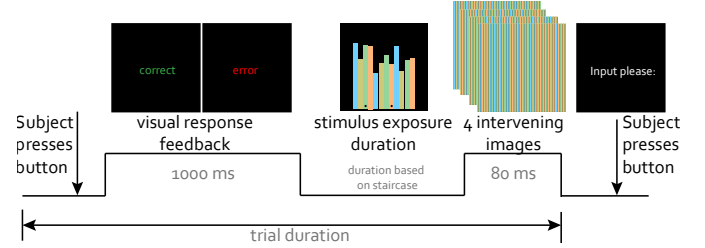

Fig. 4. Procedure for one trial in the study. After a participant's response, we show feedback on the correctness of the answer. After 1 s, the next stimulus is shown followed by four intervening images and then a black screen with a prompt to enter a response.

Table 1. Average task duration (Avg) measured in the pilot study and start time (Start) used for each staircase in our study.

	7		12		24	
	AVG	START	AVG	START	AVG	START
Bar	1344 ms	2800 ms	1557 ms	3100 ms	2517 ms	4000 ms
Donut	1594 ms	3100 ms	1482 ms	3000 ms	1555 ms	3100 ms
Radial	3162 ms	4700 ms	4305 ms	5800 ms	7455 ms	9000 ms

Next, the stimulus was shown for a certain stimulus exposure duration, followed by four intervening images. Then, participants were prompted to input an answer for the current stimulus. After finishing with one *chart type*, we asked participants if they developed any strategy while performing the task. After completing all three *chart types*, participants filled out a post-study questionnaire ranking, for each *data size*, the charts based on preference and their confidence of using the chart.

We determined the starting time for each staircase in a pilot study. Three of the authors conducted 30 trials per condition. Stimuli were shown until a participant input an answer. We calculated the average task duration and added 1500 ms to ensure a conservative starting time. Table 1 shows the results as an average over the three authors and trials as well as the starting time for each condition.

4.3 Generation of Study Stimuli

Using charts of each *data size*, participants compared two target data values. On each chart we created the larger target data value to be 250±20 fictive units of data. The smaller target data value was always 75% of the larger. We randomly generated distractors of values between 145–275 and always included at least one smaller distractor than the smallest possible target data value (<172.5) and one larger distractor than the largest possible target data value (>270) to prevent easy tasks in which one of the targets is the biggest or smallest. Based on previous studies [10,44], in which participants had to estimate the percentage difference between two targets, we restricted the difference of the two bars to a fixed value. We chose 25% as the difference because we found it was neither too big nor too small to clearly impact the results.

We drew each *chart type* using D3 within a range of 240×240 px for bar charts and 210×210 px for donut and radial bar charts. We

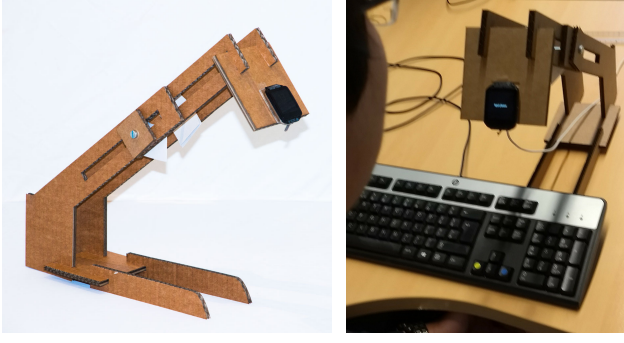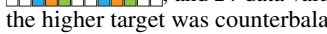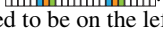Fig. 5. Left: Study setup, showing a Sony SmartWatch 3 attached to an adjustable stand. Right: A participant in front of the smartwatch with a regular keyboard on which the left and right arrow buttons are marked with a yellow and blue dot for participants' input.
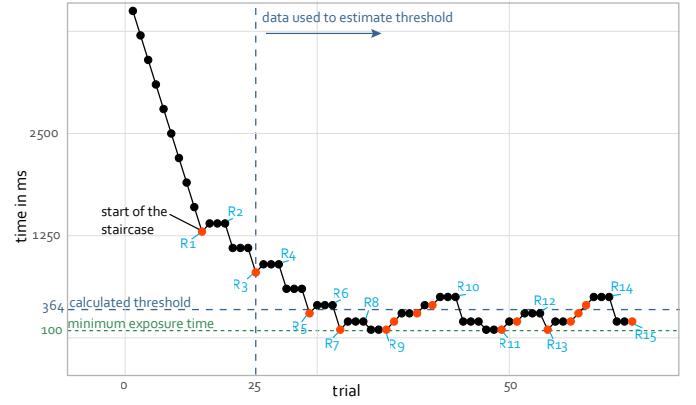

Fig. 6. Staircase from P02 for the bar chart with *data size* 24. Red trials are errors. The vertical blue line indicates the third reversal, beyond it all reversal points are averaged to calculate the threshold (horizontal blue line). Reversal points R1–R15 are labeled.

highlighted the two target data values using black dots at the bottom of each bar for the Bar ⅠⅠ, following Talbot et al.'s [44] experiment on the perception of bar charts. For the Donut ◯ we added the black dots in the center of the donut sector and for the Radial ⓖ we added it at the beginning of the bar. We considered highlighting the targets using a specific color hue but decided against this method to ensure that targets did not immediately stand out and to more closely resemble real-life situations in which participants would choose bars or sectors based on labels rather than highlights. We did not add text labels, data axes, or grid lines to avoid further distractors and ensure that we would measure comparison time.

We colored the bars and sectors with four repeating color-blind friendly isoluminant colors to ensure that the contrast between the black dot and the colored bar or sector was comparable for different colors. Opposed to other studies [10, 44], we chose to color-code the bars and sectors to ensure a closer match with the use of these charts in existing smartwatch applications. In addition, during pilot trials we found that in particular for the bar chart and radial bar chart, the task of visually tracing the bars became too difficult without a color coding.

We varied the position of the target data values and ensured a similar amount of distractors to the left and right. The two targets were placed with a distance of about 95 px between them. This means that for different *data sizes*, different amounts of distractors were placed between the two targets. We chose a distance of 95 px because we wanted to have at least two distractors between the two targets for the charts with *7* data values, following Talbot et. al's [44] finding that it is more difficult to compare separated bars than aligned ones. Possible positions for the two targets per *data size* correspond to the pairs of matching colors in the following images: *7* data values: ▢▢▢▢▢▢▢, *12* data values: ▢▢▢▢▢▢▢▢▢▢▢▢, and *24* data values: ‖‖‖‖‖‖‖‖‖‖‖‖. The position of the higher target was counterbalanced to be on the left and right. We produced the charts for each *data size* using the same data. Overall, we created 396 images to have an equal amount for each condition and enough stimuli to run a staircase procedure with ten practice trials. The stimuli were presented in a random order for each participant.

4.4 Participants

We recruited 18 participants (7 female, 11 male; 10 researchers, 8 students), with an average age of 30 years ($SD = 7.7$). Their highest degree was Bachelor (7), Master (8), and PhD (3). All participants had an CS background while most of them were either from the domain of HCI (9) or visualization (5). All participants had normal or corrected-to-normal vision and reported to have no color vision deficiency. Participants were not compensated other than with a bar of chocolate at the end.

Two participants reported to own a wrist-worn device (Fitbit and Garmin for Running). Twelve participants reported to have experience with visualizations, on average for 6.83 years ($SD = 7.12$). Participants rated their familiarity with Bar ⅠⅠ ($M = 4.89$, $SD = 0.32$), Donut ◯ ($M = 4.11$, $SD = 1.23$), and Radial ⓖ ($M = 2.28$, $SD = 1.27$) on a 5-point Likert scale (1: not familiar at all and 5: very familiar).

4.5 Apparatus

We used a Sony SmartWatch 3 with an Android Wear 2.8.0 operating system. The smartwatch has a viewable screen area of 28.73 mm × 28.73 mm and a screen resolution of 320 px × 320 px (= a pixel size of 0.089 mm). Fig. 1 (right) shows an image of the smartwatch with a Bar ⅠⅠ stimulus.

Based on the results of our pre-study (cf. Sect. 3), we attached the smartwatch to a self-designed adjustable stand at an angle of 50° (cf. Fig. 5, left). At the beginning of the study, we adjusted the stand so that the smartwatch was placed at a viewing distance of 28 cm, 20 cm height from the table surface, and roughly 90 cm from the floor. Adjustments were made while the participants were comfortably seated. We allowed participants to adjust their sitting position during the study but did not re-adjust the position of the stand because viewing distances during our pre-study also showed variability. Participants answered either *left* or *right* using the arrow keys on a keyboard placed directly in front of them to indicate which target was bigger (cf. Fig. 5, right).

We used a Dell XPS 13 laptop with Windows 10 to run a Java Program that recorded the key presses of participants, wrote a log file, determined each stimulus' exposure duration based on participants' input, and whether or not the termination criteria had been reached. The smartwatch and the laptop were connected via a Wifi hotspot and communicated using the user datagram protocol (UDP).

4.6 Measure

The measure in our study is the *time threshold* for each staircase. Given our study design, this threshold should represent ~91% correct responses for the particular combination of *chart type × data size* [16]. To compute this threshold, Gárcia-Pérez [16] recommends using the average stimulus (time in our case) of the *reversal* points in the staircase, i. e., trials in which participants oscillate between decrease-increase and increase-decrease. In particular, the author recommends to consider data after the second reversal point. Following these recommendations, for each participant and each staircase, we compute the threshold as the mean time of all reversal points after the second reversal. Fig. 6 shows an example staircase with reversal points marked.

5 STUDY RESULTS

We analyze, report, and interpret all our inferential statistics using interval estimation [12]. We report sample means of thresholds and 95% confidence intervals (CIs). We can be 95% confident that this interval includes the population mean. These results are highly representative of the plausible values of the true population mean, and the approach supports future replication efforts. We use BCa bootstrapping to construct confidence intervals (10,000 bootstrap iterations). CIs of mean differences were adjusted for multiple comparisons using Bonferroni correction [27]. We analyze the CIs using estimation techniques, i. e.,
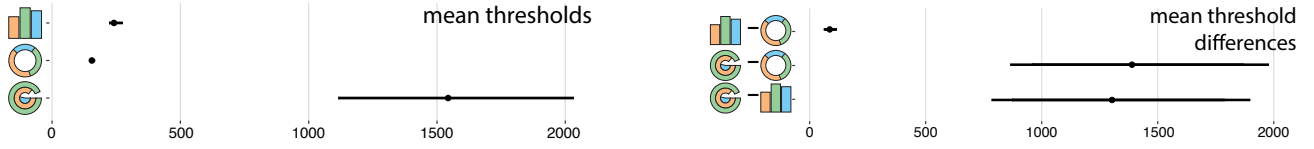
Fig. 7. Left: Average thresholds in milliseconds for each *chart type* over all *data sizes*. Right: Pair-wise comparisons between *chart types*. Error bars represent 95% Bootstrap confidence intervals (CIs) adjusted for three pairwise comparisons with Bonferroni correction.
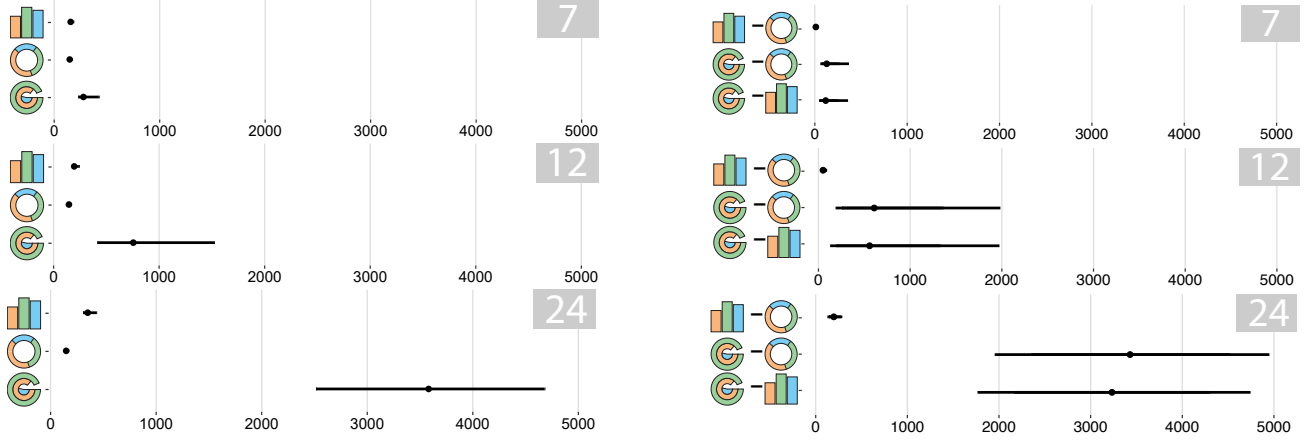

Fig. 8. Left: Average thresholds in milliseconds for each *chart type* and *data size*. Right: Pair-wise comparisons for each *chart type* and *data size*. Error bars represent 95% Bootstrap confidence intervals (CIs) adjusted for nine pairwise comparisons using Bonferroni correction.

interpreting them by providing different strength of evidence about the population mean, as recommended in the literature [3, 11, 12, 17, 21, 40]. Nonetheless, a p-value approach of our technique can be obtained following the recommendations from Krzywinski and Altman [31].

5.1 Thresholds

We collected 162 staircases from the 18 participants, for each of the 9 conditions. For each staircase we calculated the threshold as the mean of all reversals after the second reversal. Ten (out of 162) staircases terminated before the participant reached one of the two termination criteria (cf. Sect. 4.1) due to Wifi connectivity issues between the smartwatch and the laptop. Of these early terminations, eight contained at least three reversals and we were able to use these to calculate the thresholds. The remaining two staircases contained 89 and 149 completed trials, respectively, and had no reversal points because the participant had completed all (89, 149) trials without any error. In these two cases, the minimum exposure time of 100 ms was reached for the large majority of these trials, and we set the threshold to 100 ms.

Fig. 7 and Table 2 (leftmost column) summarize the CIs of the means of the three *chart types*, and of their differences of means, for all *data sizes*. They provide strong evidence that the threshold for Radial ⓖ is higher than the other two *chart types* by more than 1000 ms on average in both cases. There is strong evidence that the threshold for Bar ⓘ is higher than Donut ◯, but the difference between the two is small (86 ms). The size of the confidence intervals of the means in Fig. 7 (left) also indicates that overall, the mean threshold for Radial ⓖ may vary, whereas Donut ◯ and Bar ⓘ are more stable (i. e., the difference between the likely lower and upper bounds is small).

A closer inspection of different *data sizes* (cf. Fig. 8 and Table 2, columns 2-4) reveals that most conditions exhibit a difference between *chart types*. For the smallest *data size* (7 data values) we did not find evidence of difference between Donut ◯ and Bar ⓘ (CI touches 0; cf. Fig. 8, top left). The mean difference between Radial ⓖ and the other two *chart types* is <130 ms. (cf. Fig. 8, top right and Table 2, column 2). We consider this difference in milliseconds to be small in practice. The results across all techniques (including Radial ⓖ) are more consistent in the 7 data value condition as highlighted by the short confidence intervals obtained (cf. Fig. 8, top left).

The situation changes for larger *data sizes*. For *12* data values (cf. Fig. 8, middle left and Table 2, column 3) participants' thresholds

Table 2. Top: Mean thresholds (T) and confidence intervals (CI) across all sizes and for each *data size* given in milliseconds (p=0.05). Bottom: Pair-wise comparisons (Δ T) between *chart types* (CIs were adjusted for 3 pairwise comparisons with Bonferroni correction, p=0.0169), and for each *data size* (CIs adjusted for 9 pairwise comparisons, p=0.0056).

	ALL SIZES		7		12		24	
	T	CI	T	CI	T	CI	T	CI
Bar ⓘ	245	[221,272]	168	[155,181]	208	[187,240]	360	[307,432]
Donut ◯	159	[155,164]	157	[147,165]	158	[154,164]	162	[158,168]
Radial ⓖ	1548	[1116,2030]	286	[228,422]	766	[415,1519]	3593	[2515,4679]

	Δ T	CI	Δ T	CI	Δ T	CI	Δ T	CI
Bar-Donut	86	[60,117]	10	[-2,29]	50	[19,95]	197	[128,290]
Radial-Donut	1389	[863,1979]	129	[60,370]	608	[187,1984]	3430	[1953,4950]
Radial-Bar	1303	[783, 1899]	118	[45,360]	558	[127,1974]	3233	[1765,4745]

become worse for Radial ⓖ compared to the other two *chart types*. The differences between both charts and Radial ⓖ are close to 600 ms. The differences between Bar ⓘ and Donut ◯ is 50 ms and while the confidence interval of the difference between both techniques does not touch 0, its lower bound is only 19 ms (cf. Fig. 8, middle right). We, therefore, do not consider this as evidence for a difference between the two charts. Already in this *data size* we start seeing larger variability with Radial ⓖ, which the large CIs show (cf. Fig. 8, middle left).

For *24* data values (cf. Fig. 8, bottom left and Table 2, column 4) there is strong evidence that thresholds for Radial ⓖ are higher than for the other two *chart types* by more than 3000 ms on average. There is also evidence that Donut ◯ thresholds are lower than Bar ⓘ for this *data size*, with a difference of 197 ms (cf. Fig. 8, bottom right).

5.2 Accuracy

Although our main measure in the study is the time threshold, we report additional information on accuracy results. According to previous work [16], our study was designed to calculate thresholds when participants reached approximately 91% correct responses (error of 9-10%) for each condition. The error rates for four conditions were indeed between 9-10% (Donut ◯: 7 data values = 8%, 12 data values = 9%, 24 data values = 10%; Bar ⓘ: 7 data values = 10%). The remaining conditions, however, had somewhat higher error rates, between 16-25%

Table 3. Ranking (RK) of the three *chart types* for each *data size* (DS) for the first and second study. Top: *Chart types* participants preferred. Bottom: *Chart types* participants felt most confident with.

| | | RANKING OF CHART PREFERENCE | | | | | |
| | | Bar | | Donut | | Radial | |
DS	RK	1ST	2ND	1ST	2ND	1ST	2ND
	1	6	10	11	8	1	0
7	2	8	6	5	10	5	2
	3	4	2	2	0	12	16
	1	4	8	12	10	2	0
12	2	10	8	4	8	4	2
	3	3	2	2	0	13	16
	1	4	6	12	12	2	0
24	2	12	11	5	6	2	1
	3	2	1	1	0	14	17
		CONFIDENCE RANKING PER CHART					
	1	7	7	10	11	1	0
7	2	10	10	7	7	1	1
	3	1	1	1	0	16	17
	1	5	4	12	14	1	0
12	2	9	13	5	4	4	2
	3	4	1	1	0	13	16
	1	7	2	10	16	1	0
24	2	10	14	7	2	1	2
	3	5	1	0	0	13	16

Table 4. Average task duration (Avg) measured in the first study and start time (Start) used for each staircase in the second study.

| | 7 | | 12 | | 24 | |
	AVG	START	AVG	START	AVG	START
Bar	168 ms	1700 ms	208 ms	1700 ms	360 ms	1200 ms
Donut	157 ms	1700 ms	158 ms	1700 ms	162 ms	2300 ms
Radial	286 ms	1790 ms	766 ms	1700 ms	3593 ms	5100 ms

(Bar : 12 data values = 19%, 24 data values = 25%; Radial : 7 data values = 16%, 12 data values = 20%, 24 data values = 19%). We hypothesize that the error rate did not converge to 9-10% in these staircases because for some participants the maximum number of trials we set per staircase (150) was not enough to reach their true threshold (converge), and their times varied to a larger degree around the threshold area. This is supported by a strong correlation between variations in time (expressed by the standard deviation σ) and error rate for the different conditions (Spearman's ρ = .83, p < .01). When looking at aggregated error rates per technique (all corrections Bonferroni), we found no evidence of differences between Radial and Bar 0% [-2%,3%], and evidence of small differences between Bar and Donut 8% [5%,10%] and Radial and Donut 8% [5%,10%].

5.3 Strategies

At the end of each *chart type*, we asked participants if they applied a specific strategy to perform the task (cf. Fig. 3). In the following, we report only on strategies that multiple participants applied.

Bar : Most participants (9) reported to focus on the overall shape of the bar chart trying to estimate if the left or the right side was the high or low end of the shape. This did not always work if there were many distractors with high values on one or both sides. A second strategy participants applied (4) was to focus solely on the left or right target and try to guess just from this one bar if it was the higher bar or not.

Donut : Some participants (5) reported that they focused on the center of the screen to estimate if the left or right target was the larger element. Other participants (4), similar to the Bar , focused on only one of the targets on the left or right and tried to estimate from this if it was the larger or smaller element.

Radial : Overall, most participants (15) reported that instead of following the bar to compare the two targets, they rather looked at the gap between the beginning and end of the left (inner) target bar. Participants used the size of this gap as an indication to judge if the other bar was longer or shorter. If the gap was small the chance of this bar being the target bar were high, if the gap was large participants assumed it not to be the target. Another strategy used by participants (7) was to look at the general surrounding area on the right side of the graph. If they could spot some bars and especially a bar colored in the target's color, they chose this as the larger target, else the other one.

5.4 Post-Questionnaire

After the study, participants ranked the charts for each *data size* based on their preference and their level of confidence in performing the task correctly. Table 3 shows the average rankings for the different conditions. Overall, the Donut was preferred followed by the Bar and last the Radial for all *data sizes* (cf. Table 3, top). For confidence, it was the same ranking: Donut , Bar , and last Radial again for all *data sizes* (cf. Table 3, top).

6 EVALUATING RANDOM DIFFERENCES

We conducted a follow-up study with randomized data to minimize the effect of the strategies participants reported and to have a more diverse difference in bar heights. The study design and procedure, as well as apparatus remained the same.

We made three main changes for the follow-up study: (1) to determine the start time for each staircase, we calculated the average task duration from the first study and added ~1500 ms to ensure a conservative starting time (see Table 4); (2) through the use of an alternative network protocol (transmission control protocol, TCP) we eliminated Wifi connectivity issues and all participants finished all conditions as expected, only one participant finished 150 trials before the 15th reversal; (3) we generated charts in which the first target bar had a size between 40 and 270 data values (generated randomly) and the second between 30 and a max of $bar\_value_1 - 10$, to ensure that there was at least a ten data value difference between the two targets.

6.1 Participants

We recruited 18 new participants (7 female, 11 male; 7 researchers, 3 students), with an average age of 35 years (SD = 13). Their highest degree was High School (3), Bachelor (2), or Master (12). Participants had a background in computer science (13) (with 3 trained in HCI and 5 in visualization), marketing (2), engineering (1), secretary (1), or housewife (1). All participants had normal or corrected-to-normal vision and reported to have no color vision deficiency. Participants were not compensated other than with a bar of chocolate at the end.

Three participants reported to own a wrist-worn device (Microsoft Band and Garmin for Running). Twelve participants reported to have experience with visualizations, on average for 4.5 years (SD = 1.62). Participants rated their familiarity with Bar (M = 4.60, SD = 1.00), Donut (M = 4.28, SD = 1.13), and Radial (M = 2.33, SD = 1.65) on a 5-point Likert scale (1: not familiar at all and 5: very familiar).

6.2 Study Results

Thresholds: We collected 162 staircases in total, and computed thresholds as in the first study. One staircase terminated with less than 15 reversals, but contained at least three, so we were able to calculate the threshold. The results are shown in Fig. 9 and Table 5. The means obtained for Bar and Donut are 245 ms and 159 ms, respectively but a much higher value of 1548 ms for Radial . In Fig. 9, the non-overlapping confidence intervals provide strong evidence that Bar and Donut perform better than Radial by at least 700 ms and up to 1900 ms. However, contrary to the results obtained in the first study, in the case of completely randomized data, the overlap between the CI of Bar and Donut mean thresholds, and the overlap of their difference with 0 (cf. Fig. 9 and Fig. 10) do not provide evidence of a difference between these two *chart types*. Looking at the results for each *data size* presented in Fig. 10, however, we see a similar pattern to the results of the first study: for 7 data values, the differences between Radial and the other two *chart types* are small and increase as

Table 5. Follow-up study with randomized data. Top: Mean thresholds (T) and confidence intervals (CI) across all sizes and for each *data size* given in milliseconds (p=0.05). Bottom: Pair-wise comparisons (ΔT) between *chart types* (CIs were adjusted for 3 pairwise comparisons with Bonferroni correction, p=0.016), and for each *data size* (CIs adjusted for 9 pairwise comparisons, p=0.0056).

	ALL SIZES		7		12		24	
	T	CI	T	CI	T	CI	T	CI
Bar 📊	285	[252,333]	181	[167,209]	232	[204,276]	443	[376,538]
Donut ⭕	216	[180,337]	178	[162,231]	202	[175,262]	267	[197,507]
Radial 🌀	1772	[1513,2017]	560	[413,784]	853	[696,1151]	3905	[3273,4439]
	ΔT	CI	ΔT	CI	ΔT	CI	ΔT	CI
Bar-Donut	69	[-6,114]	2	[-22,24]	30	[-14,83]	176	[-28,296]
Radial-Donut	1556	[1257,1814]	381	[199,696]	650	[457,1081]	3638	[2808,4295]
Radial-Bar	1487	[1173, 1756]	379	[195,717]	620	[428,1021]	3461	[2598,4201]

the number of data values increases. However, contrary to the first study, this study with randomized data does not provide any evidence to suggest a difference between Bar 📊 and Donut ⭕.

Accuracy: Error rates were higher than in the first study and not between 9-10%. Donut ⭕ had the lowest error rates (7 data values = 13%, 12 data values = 13%, 24 data values = 18%). Followed by Bar 📊 (7 data values = 16%, 12 data values = 23%, 24 data values = 29%) and last Radial 🌀 (7 data values = 28%, 12 data values = 28%, 24 data values = 31%). When looking at aggregated error rates per technique (all corrections Bonferroni), we found small evidence of differences between Radial 🌀 and Bar 📊 5% [3%,8%], and Bar 📊 and Donut ⭕ 7% [5%,9%], but a somewhat higher difference between Radial 🌀 and Donut ⭕ 13% [10%,15%].

Strategies: Overall, participants reported fewer strategies in this study. Bar 📊: Participants (6) chose which target was smaller or larger by focusing on the center of the chart and estimating based on the overall shape of the bar chart if there were many large or small bars on one side. Else participants (4) reported that if one of the targets was very small or large, they made their decision based on this one target. Donut ⭕: The strategy most participants (7) used, was to focus on the center of the chart and estimate which target is larger. Radial 🌀: Most participants (11) told us that they looked a the inner bar and tried to estimate from this if it was the smaller or larger one.

Post-Questionnaire: As in the first study, we asked participants to rank the three charts based on preference (cf. Table 3, top) and confidence (cf. Table 3, bottom). Similar to the first study, the Donut ⭕ was preferred most, followed by the Bar 📊, and last the Radial 🌀. The only exception is the 7 data value condition; here more participants ranked the Bar 📊 as their number one and then the Donut ⭕. This is the only difference to the first study. For confidence, the ranking was the same as for the first study: first Donut ⭕, then Bar 📊, and last Radial 🌀.

7 DISCUSSION

We set out to assess how quickly people can perform a simple data comparison task for small-scale visualizations on a smartwatch. Our study results give first evidence to answer this question. For a small number of data values, participants could estimate if one data value is higher than the other one for all three *chart types* (with 70–92% accuracy) in times between 160 ms–560 ms. For bar charts and donut charts in particular, they could reliably perform this task when dealing with 24 data values, on average in ~450 ms using bar charts, and impressively in ~270 ms using donut charts. These thresholds are slightly above the 200 ms that the scene perception literature considers as a "glance" [34]. Even the threshold of the slowest technique with large data values (radial bar chart, with ~3900 ms) was as low as or lower than times reported in smartwatch usage studies (cf. Sect. 2.1). Based on the study results, we discuss design considerations for glanceable smartwatch visualizations, the limitations of our study and potential further research directions.

7.1 Reflecting on Study Design and Results

Where the differences between the three charts come from is an interesting question. Our design of the visual stimuli influenced our results in two ways. First, we used a minimum number of 7 data values (with at least 5 distractors) for the study stimuli, ruling out that participants could memorize the individual data items for a single chart [52]. Second, in terms of visual processing, we speculate that different visual routines [47] were involved. In particular, the *chart type* varied how participants had to associate the location of the black dots with the values to be compared. In our donut design, dots were placed at a 50% reference point. In bar and radial charts, they were placed at the beginning of the bar (i. e., aligned to the axis), to mimic label placement and to follow previous studies in the visualization literature [44]. For these two charts a (slower) spatial shift of attention may have been necessary [38] because participants had to trace a much longer distance. If the dots had been placed at the top of the bars in both types of bar charts the results may have been much faster, in particular for the radial bar chart. However, bar chart labels are more common at the bottom of the bars. The dot placement likely plays a role in the ease and speed of association, and it is possibly related to previous findings [44] on the dot location influencing the correctness of bar height estimation (dots in the middle made comparisons more correct).

In addition to the dot location, the variance of colors and lengths of the non-target bars likely also slowed the search for the endpoint [25]. Generally, conditions that demanded more attentional shifts, involved more complex target-to-endpoint (dot to value) association, yielded more imprecise spatial selection, and were more susceptible to surrounding variance resulted in higher response times. In addition, the strategies participants described suggest that participants did not necessarily perform the task on both marked targets. Instead they regularly performed an estimation of the size of one target and then made a guess, without consulting the other target. Given these strategies, the low thresholds our participants reached, in some cases hitting the minimum threshold of 100 ms without errors, are perhaps not surprising.

Yet, this fast strategy can also be applicable in real life smartwatch use cases, for example, quickly glancing at one's physical activity on two different days (Monday vs. Wednesday) or to an average bar drawn on the side of daily activity data.

7.2 Design Considerations

Situations in which visualizations need to be looked at quickly are common on smartwatches. For example, visualizations can convey several data values as part of a notification. Our study results showed that the heights of individual bars or donut sectors up to 24 data values can be assessed within a few hundred milliseconds. Radial bars up to 7 data values can also be assessed quickly. Radial bar charts of higher data values had much larger thresholds and varied widely across participants. Furthermore, they were the least preferred and gathered the lowest confidence scores. Another disadvantage of radial bar charts is the available bar width. Due to its encoding, the bars in radial bar charts are roughly half the size of a bar chart for the same number of items, making any discrimination task more challenging.

The results acquired from our studies also highlight that both bar charts and donut charts provide similar results. For 7 and 12 data values, the differences are under 100 ms while for 24 it is under 200 ms. These relatively small time differences lead us to believe that when creating visualizations, designers can use them almost interchangeably. Given the similar performance of donut and pie charts found in previous studies [41], pie charts may also be possible candidates.

Nevertheless, the stable performance of donut charts across all *data sizes* (going from 7 to 24 data values only increases the threshold by <100 ms on average) may indicate that this visualization could scale to more than 24 data values for the particular task. If designers are considering even larger *data sizes*, then this visualization could be considered—keeping in mind the readability of the chart's labels and segment color may be impacted with more data values.

It is not straightforward to speculate on the generalizability of our results to larger display sizes. The charts we chose, their design (e. g., no labels, axes, grid lines, tick marks), as well as the number of data
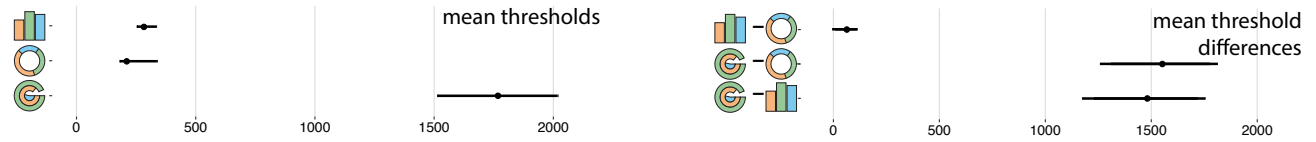
Fig. 9. Left: Average thresholds in milliseconds for each *chart type* over all *data sizes*. Right: Pair-wise comparisons between *chart types*. Error bars represent 95% Bootstrap confidence intervals (CIs) adjusted for three pairwise comparisons with Bonferroni correction.
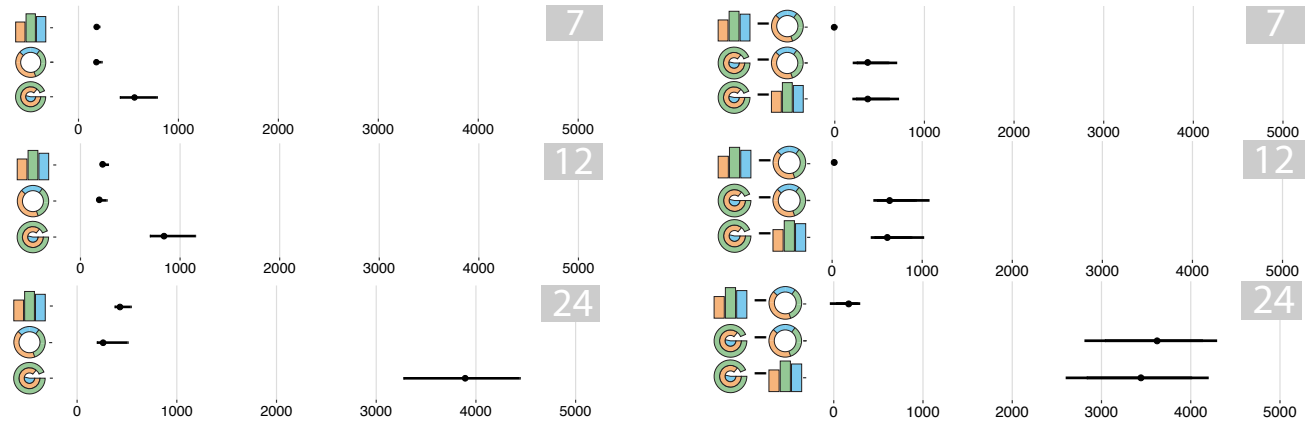

Fig. 10. Left: Average thresholds in milliseconds for each *chart type* and *data size*. Right: Pair-wise comparisons for each *chart type* and *data size*. Error bars represent 95% Bootstrap confidence intervals (CIs) adjusted for nine pairwise comparisons with Bonferroni correction.

items we tested (7, 12, 24) were influenced by our smartwatch usage scenario. It is even possible that our small display size is beneficial for this type of task, as the entire visualization covers on average a small visual angle (6°, with targets being in the central vision field) and can be seen with little to no eye movement. Nevertheless, further research is required to verify this. In addition, the task that requires performing comparisons, may be more important to smartwatch or smartphone usage, during which people often quickly glance at their devices.

7.3 Limitations and Future Work

The low thresholds we calculated for some conditions lead us to conclude that actual thresholds for this particular task could be lower than our limit of 100 ms. It is important, though, to note that our task was simple because it had a fairly low search time, as the rough position of the dots did not vary much between trials. More complex tasks in which viewers need to read and search for labels probably have higher time thresholds. Even with the randomized data for the two targets, we still had low thresholds, confirming that our task was simple.

Different familiarity levels with the three *chart types* and the relatively young age group of our participants could have affected our results. On the one hand, all three *chart types* and the task we studied are fairly simple and participants conducted practice trials. On the other hand, all of our participants in the first study and most in the second study, had a computer science background, and many of them were from the HCI or visualization domain. Therefore, our results are likely to be a "best case" and thresholds might be higher for a population with less comfort or experience reading charts.

In addition, our thresholds should also be considered as a lower bound because we used a static watch. In practice, we do not keep our arm quite steady especially while we are walking or running. An interesting direction for future work is to study glanceability in-situ while people are moving or engaged with social activities (e. g., chatting, running or eating with other people).

Furthermore, while we studied only one simple data comparison task, there are other tasks that can be performed on smartwatches such as detecting trends, searching for extreme values, etc. It would be interesting to study different types of tasks and determine their specific thresholds. Another interesting follow-up would be to investigate when the threshold for the bar chart and donut chart would diverge when the number of data values shown increases or decreases.

8 CONCLUSION

We present two perceptual studies on smartwatches that try to determine how quickly people can perform a simple data comparison task on these small displays. In particular, we compared three *chart types* that are common in smartwatch applications: bar charts, donut charts, and radial bar charts. For each *chart type* we also compared three *data sizes*: 7, 12, and 24 data values. In our first study that required the comparison of two visual marks, which differed by 25%, participants were able to reach time thresholds as low as 160–210 ms for 7 and 12 data values for bar charts and donut charts. These techniques also scaled well to 24 data values, average times around 360 ms and 163 ms, respectively. Whereas the performance of radial bar charts dropped considerably from 286 ms for 7 data values to 766 ms for 12 and to 3600 ms for 24 data values. The same trends were also present in our second study, in which differences between marks were randomly generated. However, in the second study the threshold values went up by around 50 ms on average for bar charts and donut charts, and more substantially for radial bar charts by approximately 200 ms. Because of their overall poor performance, radial bar charts seem unsuitable for smartwatch applications that require quick comparisons if the number of data values is higher than seven.

Our work opens up several directions for future research on smartwatch visualizations, which become more prevalent to convey quantitative information. For example, the values of average viewing angle and distances when reading smartwatches, obtained from our pre-study and used in our main studies, can serve as a basis for further controlled experiments on smartwatches.

REFERENCES

[1] F. Beck, T. Blascheck, T. Ertl, and D. Weiskopf. Exploring word-sized graphics for visualizing eye tracking data within transcribed experiment recordings. In *Proceedings of the Workshop on Eye Tracking and Visualization (ETVIS)*, pp. 1–5, 2015.

[2] F. Beck and D. Weiskopf. Word-sized graphics for scientific texts. *IEEE Transactions on Visualization and Computer Graphics*, 23(6):1576–1587, 2017. doi: 10.1109/TVCG.2017.2674958

[3] L. Besançon and P. Dragicevic. The significant difference between p-values and confidence intervals. In *Proceedings of the Conference on l'Interaction Homme-Machine (IHM)*, pp. 53–62. ACM, 2017.

[4] T. Blascheck, A. Bezerianos, L. Besançon, B. Lee, and P. Isenberg. Preparing for perceptual studies: Position and orientation of wrist-worn smartwatches for reading tasks. In *Proceedings of the Workshop on Data Visualization on Mobile Devices held at ACM CHI*, pp. 1–6, 2018.

[5] R. Borgo, J. Kehrer, D. Chung, E. Maguire, R. Laramee, H. Hauser, M. Ward, and M. Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. In *Proceedings of Eurographics – State of the Art Reports*, pp. 39–63. The Eurographics Association, 2013. doi: 10.2312/conf/EG2013/stars/039-063

[6] U. Brandes. Visualization for visual analytics: Micro-visualization, abstraction, and physical appeal. In *Proceedings of the Pacific Visualization Symposium (PacificVis)*, pp. 352–353. IEEE Computer Society Press, 2014. doi: 10.1109/PacificVis.2014.67

[7] U. Brandes, B. Nick, B. Rockstroh, and A. Steffen. Gestaltlines. *Computer Graphics Forum*, 32(3):171–180, 2013. doi: 10.1111/cgf.12104

[8] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann Publishers, 1st ed., 1999.

[9] Y. Chen. Visualizing large time-series data on very small screens. In *Proceedings of EuroVis – Short Papers*, pp. 37–41. The Eurographics Association, 2017. doi: 10.2312/eurovisshort.20171130

[10] W. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *American Statistical Association*, 79(387):531–554, 1984. doi: 10.2307/2288400

[11] G. Cumming. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Routledge, 1st ed., 2013. doi: 10.1111/j.1751-5823.2012.00187_26.x

[12] P. Dragicevic. Fair statistical communication in HCI. In J. Robertson and M. Kaptein, eds., *Modern Statistical Methods for HCI*, pp. 291–330. Springer, 2016. doi: 10.1007/978-3-319-26633-6_13

[13] A. Esteves, E. Velloso, A. Bulling, and H. Gellersen. Orbits: Gaze interaction for smart watches using smooth pursuit eye movements. In *Proceedings of Symposium on User Interface Software & Technology (UIST)*, pp. 457–466. ACM, 2015. doi: 10.1145/2807442.2807499

[14] J. Fuchs. *Glyph Design for Temporal and Multi-Dimensional Data: Design Considerations and Evaluation*. PhD thesis, Universität Konstanz, 2015.

[15] J. Fuchs, P. Isenberg, A. Bezerianos, and D. Keim. A systematic review of experimental studies on data glyphs. *IEEE Transactions on Visualization and Computer Graphics*, 23(7):1863–1879, 2017. doi: 10.1109/TVCG.2016.2549018

[16] M. García-Pérez. Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties. *Vision Research*, 38(12):1861–1881, 1998. doi: 10.1016/S0042-6989(97)00340-4

[17] G. Gigerenzer. Mindless statistics. *The Journal of Socio-Economics*, 33(5):587–606, 2004. doi: 10.1016/j.socec.2004.09.033

[18] P. Goffin. *An Exploration of Word-Scale Visualizations for Text Documents*. PhD thesis, Université Paris Saclay, 2016.

[19] P. Goffin, J. Boy, W. Willett, and P. Isenberg. An exploratory study of word-scale graphics in data-rich text documents. *IEEE Transactions on Visualization and Computer Graphics*, 23(10):2275–2287, 2017. doi: 10.1109/TVCG.2016.2618797

[20] P. Goffin, W. Willett, J.-D. Fekete, and P. Isenberg. Exploring the placement and design of word-scale visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2291–2300, 2014. doi: 10.1109/TVCG.2014.2346435

[21] S. N. Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995–1004, 1999. doi: 10.7326/0003-4819-130-12-199906150-00008

[22] R. Gouveia, F. Pereira, E. Karapanos, S. A. Munson, and M. Hassenzahl. Exploring the design space of glanceable feedback for physical activity trackers. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 144–155. ACM, 2016. doi: 10.1145/2971648.2971754

[23] M. Greene and A. Oliva. The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4):464–72, 2009. doi: 10.1111/j.1467-9280.2009.02316.x

[24] B. Greenhill, M. Ward, and A. Sacks. The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science*, 55(4):991–1002, 2011. doi: 10.1111/j.1540-5907.2011.00525.x

[25] S. Haroz and D. Whitney. How capacity limits of attention influence information visualization effectiveness. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2402–2410, 2012. doi: 10.1109/TVCG.2012.233

[26] J. Heer, N. Kong, and M. Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pp. 1303–1312. ACM, 2009. doi: 10.1145/1518701.1518897

[27] J. J. Higgins. *Introduction to Modern Nonparametric Statistics*. Thomson Learning, 1st ed., 2004.

[28] T. Horak, S. K. Badam, N. Elmqvist, and R. Dachselt. When David meets Goliath: Combining smartwatches with a large vertical display for visual data exploration. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pp. 19:1–19:13. ACM, 2018. doi: 10.1145/3173574.3173593

[29] W. Javed, B. McDonnel, and N. Elmqvist. Graphical perception of multiple time series. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):927–934, 2010. doi: 10.1109/TVCG.2010.162

[30] F. Kingdom and N. Prins. *Psychophysics: A Practical Introduction*. Elsevier Science BV, 1st ed., 2010.

[31] M. Krzywinski and N. Altman. Points of significance: Error bars. *Nature Methods*, 10(10):921–922, 2013. doi: 10.1038/nmeth.2659

[32] C. Min, S. Kang, C. Yoo, J. Cha, S. Choi, Y. Oh, and J. Song. Exploring current practices for battery use and management of smartwatches. In *Proceedings of the International Symposium on Wearable Computers (ISWC)*, pp. 11–18. ACM, 2015. doi: 10.1145/2802083.2802085

[33] I. Oakley, D. Lee, M. R. Islam, and A. Esteves. Beats: Tapping gestures for smart watches. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pp. 1237–1246. ACM, 2015. doi: 10.1145/2702123.2702226

[34] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. In S. Martinez-Conde, S. Macknik, L. Martinez, J.-M. Alonso, and P. Tse, eds., *Visual Perception*, pp. 23 – 36. Elsevier Science BV, 2006. doi: 10.1016/S0079-6123(06)55002-2

[35] J. Parnow. Micro visualizations: How can micro visualizations enhance text comprehension, memorability, and exploitation. Master's thesis, Potsdam University of Applied Sciences, 2015.

[36] C. Perin, R. Vuillemot, and J.-D. Fekete. SoccerStories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2506–2515, 2013. doi: 10.1109/TVCG.2013.192

[37] S. Pizza, B. Brown, D. McMillan, and A. Lampinen. Smartwatch in vivo. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pp. 5456–5469. ACM, 2016. doi: 10.1145/2858036.2858522

[38] M. I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1):3–25, 1980. doi: 10.1080/00335558008248231

[39] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *Proceedings of the Conference on Information Visualization (InfoVis)*, pp. 173–180. IEEE Computer Society Press, 2005. doi: 10.1109/INFVIS.2005.1532144

[40] F. L. Schmidt and J. E. Hunter. Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. Harlow, S. Mulaik, and J. Steiger, eds., *What If There Were No Significance Tests?*, pp. 37–64. Lawrence Erlbaum Associates, 1997.

[41] D. Skau and R. Kosara. Arcs, angles, or areas: Individual data encodings in pie and donut charts. *Computer Graphics Forum*, 35(3):121–130, 2016. doi: 10.1111/cgf.12888

[42] M. Stone. In color perception, size matters. *IEEE Computer Graphics and Applications*, 32(2):8–13, 2012. doi: 10.1109/MCG.2012.37

[43] J. Talbot, J. Gerth, and P. Hanrahan. An empirical model of slope ratio comparisons. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2613–2620, 2012. doi: 10.1109/TVCG.2014.2346320

[44] J. Talbot, V. Setlur, and A. Anand. Four experiments on the perception of bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2152–2160, 2014. doi: 10.1109/TVCG.2014.2346320

[45] The NPD Group, Inc. Smartwatch ownership expected to increase nearly 60 percent into 2019. Online Press Release, https://www.npd.com/wps/portal/npd/us/news/press-releases/2017/us-smartwatch-ownership-expected-to-increase-nearly-60-percent-into-2019, 2017. Accessed Mar 26, 2018.

[46] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1st ed., 2001.

[47] S. Ullman. Visual routines. *Cognition*, 18(1):97–159, 1984. doi: 10.

1016/0010-0277(84)90023-4

[48] A. Visuri, Z. Sarsenbayeva, N. van Berkel, J. Goncalves, R. Rawassizadeh, V. Kostakos, and D. Ferreira. Quantifying sources and types of smartwatch usage sessions. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, pp. 3569–3581. ACM, 2017. doi: 10.1145/3025453.3025817

[49] M. Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3/4):194–210, 2002. doi: 10.1057/palgrave.ivs.9500025

[50] M. Ward. Multivariate data glyphs: Principles and practice. In 1st, ed., *Handbook of Data Visualization*, pp. 179–198. Springer, 2008.

[51] W. Willett, J. Heer, and M. Agrawala. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1129–1136, 2007. doi: 10.1109/TVCG.2007.70589

[52] W. Zhang and S. J. Luck. Discrete fixed-resolution representations in visual working memory. *Nature*, 453(7192):233–235, 2008. doi: 10.1038/nature06860