# Studies of Part-to-Whole Glanceable Visualizations on Smartwatch Faces

Tanja Blascheck, Lonni Besançon, Anastasia Bezerianos, Bongshin Lee, Alaul Islam, Tingying He, Petra Isenberg

# Studies of Part-to-Whole Glanceable Visualizations on Smartwatch Faces

Tanja Blascheck*
University of Stuttgart, Stuttgart, Germany

Lonni Besançon†
Linköping University, Sweden & Monash University, Australia

Anastasia Bezerianos‡
Université Paris Saclay, CNRS, Inria, France

Bongshin Lee§
Microsoft Research, USA

Alaul Islam, Tingying He & Petra Isenberg¶
Université Paris Saclay, CNRS, Inria, France


Study 1 Study 2 Study 3

Figure 1: Stimuli from our three studies showing proportion representations related to goals for calories burned, step count, and active minutes as bar charts, radial bar charts, and text (Study 1 only). Left: Stimuli from Study 1 with a digital time representation. Middle: Stimuli from Study 2 with an additional analog time representation. Right: Stimuli from Study 3 with a different color-coding.

## Abstract

We present three studies that investigate the effectiveness of multiple glanceable part-to-whole proportion representations on smartwatch faces. Our goal was to understand how quickly and accurately people can make judgments about their progress toward multiple goals displayed in a small space. We designed our three studies with increasing external validity. The first study compared bar charts, radial bar charts, and text representations—shown with a digital time display. The second study added an analog time dial as a distractor to increase the complexity of the watch face. To emulate realistic viewing conditions, the third study investigated the effect of viewing angles. In Study 1 bar and radial bar charts outperformed text representations, in Study 2 adding an analog time dial as a distractor did not affect task performance, and in Study 3 only the most extreme angle led to some performance decrease. Supplementary material is available at `https://osf.io/ad2z7/`.

**Index Terms:** Human-centered computing—Visualization—Visualization design and evaluation methods

## 1 Introduction

Smartwatch technology has evolved drastically: smartwatches now are smart displays that collect and retrieve data and show this data in the form of watch face dashboards. Typically, smartwatch faces contain a representation of time as well as multiple watch complications—a term from horology that describes graphical features on watches that represent information other than time [51]. Complications expose a wide range of data, including step count, active minutes, or weather information. Watch complications are often only around 1 cm×1 cm large. Typical complications contain short texts (e. g., date, temperature), icons (e. g., email, weather condition), simple visualizations (e. g., bar charts, pie charts, radial bar charts, line charts), or combinations of these [31].

Smartwatches have a unique usage context for visualization: they are often used "on the go," with average usage times of 6.7 seconds [43]. Such a short time allows for quick *peeks* or *glances*, for example, to check the time. However, it is unclear how much additional information can be seen "at a glance" and which representation type can communicate information most effectively. Previous work [10] suggests that different representation types can be read at different speeds but did not study the watch face context with its typical representation types, smaller visualization sizes, and multiple co-located representations. Therefore, we set out to investigate how quickly multiple visualizations can be read on watch faces to establish time thresholds for different representation types.

We focused our work around one of the main tasks people perform with smartwatches: monitoring their progress towards their self-set goals (e. g., 10,000 steps a day) [1]. This monitoring requires seeing the proportion of a current value concerning this goal, either as a percentage `Text70%` or as a visualization .

We conducted three perceptual studies to assess how quickly and accurately people can read common proportion representations on smartwatch faces. In each study, we increased external validity, to gain an in-depth and more practical understanding of how people read proportions *at a glance* (Fig. 1). We found that both `Bar` and `Radial` representations were faster and less error-prone than a `Text70%`. `Bar` was slightly better than `Radial` both with a digital and digital+analog time display. Overall, adding an analog time display had only a negligible effect on task performance. From the last study, we have first results that the watch viewing angle might affect performance when reading `Radial` charts.

In summary, this work contributes to understanding a nascent research topic—small-scale visualizations for smartwatches. Drawn from the in-depth reflections on the results, our work also discusses practical design implications and directions for future research.

## 2 Related Work

Proportion representations are common on smartwatches, especially for the support of part-to-whole judgments, such as a comparison of progress towards a goal (for questions such as "how much of my step count goal have I achieved today?"). Part-to-part judgments are less commonly supported on smartwatches; we rarely see visualizations that compare, for example, how much time people spent walking, sitting, or sleeping. Therefore, we focus on part-to-whole representations, and especially rectilinear and circular constructions, which we found to be most common (see Sect. 3.2). Here, we review related work focusing on the differences between the two representations. In addition, we discuss the real-world use of smartwatches, and review studies that have investigated visualizations on smartwatches.

*e-mail: research@blascheck.eu
†e-mail: lonni.besancon@gmail.com
‡e-mail: anab@lri.fr
§e-mail: bongshin@microsoft.com
¶e-mail: mohammad-alaul.islam|tingying.he| petra.isenberg|@inria.fr

### 2.1 Visualization of Proportions

In psychology, researchers investigated proportion judgments for two subdivisions (a part and a whole), similar to our goal: Simkin and Hastie [45] found pie charts to be as accurate as stacked bars, whereas other researchers [18, 21, 30] found pie charts to be more accurate than stacked bars. Even though two studies [30, 45] found that completion time was faster for stacked bars in comparison to pie charts, Eells' [21] participants answered more questions with pie charts than with bar charts in five minutes. As such, this evidence does not allow us to predict performance on smartwatches.

Part-to-whole judgments have also been extensively studied in visualization [34–38, 46, 53]. All of these studies included pie charts and their variations and some studies [36, 38] in addition compared pie charts to other charts. Kosara and Ziemkiewicz [38] compared pie charts, donut charts, stacked bar charts, and square pie charts. Participants achieved the highest accuracy with the square pie charts and the lowest accuracy with the stacked bar charts. Donut and pie charts were not significantly different from each another. Later, Skau and Kosara [46] investigated the effect of changing the inner radii (0, 20, 40, 60, 80, and 97%) of a donut chart on performance, however, only the thin donut chart with an inner radius of 97% was slightly less accurate than the rest, a result confirmed by Cai et al. [12]. Kosara [36] also compared five different part-to-whole representations composed of five parts each, asking participants to estimate the percentage of the largest or median part. The stacked bar chart was slightly faster and had similar accuracy.

The related work done so far showed that pie charts are more or equally accurate as stacked bar charts whereas stacked bar charts tend to have faster completion times. Our work adds to this line of research by investigating two novel contexts: a) judgments of small-scale proportion visualizations on smartwatches—previous work [18, 21, 36, 45] rendered pie charts at sizes of 45 mm–160 mm in diameter and the stacked bar charts between 50 mm–200 mm in length. Our radial bar chart has a diameter of 8 mm and the stacked bar chart is 8 mm long; and b) judgment tasks that require reading multiple co-located proportion visualizations at the same time.

### 2.2 Real-world Use of Smartwatches

Smartwatches have unique usage contexts for visualization research. Studies have shown that peoplelook at a smartwatch on average for only 5–7 sec [24, 25, 41, 43, 50]. Such short *peeks* or *glances* allow reading the time (avg. 1.9 sec [43]) but it is unclear how much other data can be taken in at a glance from a watch face.

To understand what people want to learn from their personal data, Choe et al. [16] analyzed 30 videos during which quantified selfers presented the insights gained from their data. The researchers found that quantified selfers engaged in many insight-generating data analysis tasks (such as looking at trends, correlations, or distributions) to self-reflect. Amini et al. [1] conducted interviews with 10 participants to elicit needs for the exploration of health and fitness data, especially while people are moving. The most common insights were related to reading single or multiple values, estimating the progress towards a goal, comparing one or multiple measures with other people's, and being motivated. This work supports our motivation to focus on the comparison of multiple proportions. Based on their findings, Amini et al. also asked nine graphics designers to sketch representations for each insight type. Especially for the goal-based insight category, most sketches used visualizations instead of text. The visualizations were either donut and pie charts or space-filling shapes such as a single bar or icon, similar to the representations we use in our experiment.

### 2.3 Studies of Visualizations on Smartwatches

Only recently have researchers started to conduct dedicated research on visualizations for smartwatches. Some of this research targets novel types of representations such as Chen's [14] time-series, Suciu and Larsen's [49] time spiral, or Neshati et al.'s [42] compressed line charts. Other types of research investigated existing visualizations on smartwatches: Aravind et al. [2], for example, asked which types of visualizations people prefer for sleep data. Islam et al. [31] studied which type and how much information people displayed on their watch face. This study result is part of our motivation to test tasks that involve multiple co-located visualizations.

A third category of work concerns perception studies such as ours. Heer et al. [28] provide a first indication that studying visualizations under different sizes and viewing conditions are important, because of the unexpectedly poor performance of filled line charts at small vertical heights. Healey et al. [27] conducted more basic perceptual experiments with size encodings and saw fewer errors and shorter response times for larger-sized stimuli. Cai et al. [12] tested donut charts of different sizes, approximately 3–9× larger than ours, and found no evidence of an effect of size on the accuracy or speed of proportion estimation tasks. Most closely related to our work is a prior study by Blascheck et al. [10] that used a similar study methodology on a smartwatch. The authors investigated three different charts (bar, donut, and radial) with an increasing number of data points shown full-screen on a smartwatch. The researchers' goal was to find minimal perception time thresholds for a simple data comparison task: finding the larger of two marked elements. Compared to that previous work, our present paper studies a completely different task: the perception of *percentages* presented *simultaneously* in three co-located visualizations (Fig. 1). Furthermore, we consider more realistic watch faces, providing insights on how to improve the design of visualizations for them. The previous work [10] used only one visualization, filling the entire smartwatch screen.

While these perception studies are valuable first steps to understanding the effects of small-scale visualization in a mobile context, they do not allow us to predict the performance of the task and setup we test in the following previous tasks.

### 2.4 Summary

Previous work inspired our endeavor to continue comparing rectilinear and circular constructions, which have been extensively studied but are common representation types on smartwatches. Because smartwatch faces have a unique style using both digital and analog time representations as well as displaying multiple data types, we took this as inspiration and compared how this affects glanceability.

## 3 Evaluating Proportion Vis for Watch Faces

In this section, we discuss our overall research questions, a pre-study we conducted to ground our decisions for the follow-up studies, and the study design shared by these three studies. Documents, data, code for the statistical analysis, detailed results, and stimuli are available in the supplemental material (`https://osf.io/ad2z7/`).

### 3.1 Research Questions

When on the go people often monitor their progress and check whether they have reached their self-set goal (e. g., 10,000 steps a day) [1]. We wanted to investigate if people can check progress towards different goals at a glance on a watch face when presented with multiple proportion representations simultaneously. Specifically, our research questions (RQ) are: (RQ1) What are common watch face representations to represent proportions (Sect. 3.2)? (RQ2) Which part-to-whole proportion representation has the lowest time threshold and highest accuracy for multiple smartwatch complications (Sect. 4–6)? (RQ3) How does the density and complexity of a watch face affect task performance (Sect. 4–6)? (RQ4) How does the viewing angle of a smartwatch affect task performance (Sect. 6)?

### 3.2 Pre-Study: Common Proportion Representations

To choose which proportion representations to evaluate, we first systematically captured their variations on smartwatches from the

"most popular" watch faces listed on the Facer Android website [40]. Facer is a popular watch face creation and delivery app, website, and community that features thousands of watch faces in 22 categories. We downloaded the top 100 watch faces for four weeks: out of 400 watch faces, 184 were unique. One author conducted the first coding on the representation type of time, the number of complications, the data types shown, and the data types depicted as a proportion representation. Another author later checked the coding.

From the 184 unique watch faces coded, 48.4% use a digital display of time, 28.2% an analog display, and 23.4% a combination of both (hybrid). The median number of complications shown on the watch face was 4 (min: 0, max: 16). We found 24 different types of data represented next to time and date. Of these, we considered seven (29.1%) as representing a proportion (calories, distance, heart rate, humidity, phone & watch battery, and steps). We further distinguished between *real* and *derived* proportions. Real proportions are humidity, phone, and watch battery. Derived proportions are derived from a personally set maximum, for example, 10,000 steps or a maximum heart rate of 220 bpm. The derived proportions are calories, distance, heart rate, and steps.

Next, we looked at the 153 watch faces that had at least one data type using a proportion representation. On average 1 (max: 4) proportion data type was represented on these watch faces. Further exploring those watch faces that showed more than one data type as a proportion (34.0%), we found that common data types shown together were: watch battery and steps (14.3%); watch battery, steps, and heart rate (5.2%); watch battery and phone battery (5.2%); watch battery, steps, and phone battery (4.5%); watch battery, steps, phone battery, and distance (1.3%); watch battery, steps, and calories (0.7%); watch battery and humidity (0.7%); watch battery, phone battery, and heart rate (0.7%); steps and humidity (0.7%); watch battery, steps, and humidity (0.7%).

We then identified two data dimensions to describe the proportion representations we had captured: the category (e. g., step count, battery level) and the proportion value (e. g., 7,283 steps, 37%) of the data. For these representations, we derived a simple design space of proportion representations. Building on Spence and Krizel's [47] distinction of *integrated* versus *separated* proportion representations, we used Bertin's [5] principal types of construction for diagrams with two data dimensions to distinguish categories.

Bertin distinguishes six types of construction, four of which lead to Spence and Krizel's *separated* proportion representations: orthogonal construction (e. g.,  ), rectilinear elevation (e. g.,  ), polar construction (e. g.,  ), and circular elevation (e. g.,  ). In all these forms of construction, the whole is not explicitly represented. While it is possible to draw a whole next to the parts (e. g., one large bar that represents the whole next to all bars representing parts) this type of construction is unusual and would require careful design to be understandable to the broad population targeted with smartwatch displays. Therefore, separated displays are not designed to show proportions easily and we did not consider them further.

This exclusion left us with Bertin's final two types of construction that show *integrated* constructions. In integrated constructions, proportions form a whole and both the whole and the parts are immediately accessible to the viewer. The proportion representations in our survey show the data either as a continuous  representation (75.3%) or as a discrete  representation (24.7%).

***Rectilinear construction.*** Proportions are juxtaposed along one linear dimension. This type of construction appeared often in our survey, with the stacked bar chart  being a common representative. We found that 25.3% of the proportion representations in our survey showed a bar chart variation. Specific representations were continuous bars (10.1%)  , discrete bars (7.0%)  , and bars within icons (8.2%)  .

***Circular construction.*** Proportions can be represented as angles  or lengths on a circumference. This construction type was the most common in our survey (74.7%). We found gauges (27.9%)  , arcs (23.5%)  , donuts (13.2%)  , sliding ranges (7.6%)  , and pie charts (2.5%)  . The arc and donut chart representations were both continuous (14.6%  , 4.4%  ) and discrete (8.9%  , 8.8%  ).

Table 1: Counts for proportion representations by data type on the coded watch faces, grouped by the type of construction—rectilinear and circular. Color coding shows the amount normalized to the max (26). Legend: bat. = battery, disc. = discrete.

| | rectilinear | | | circular | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | bar | disc. bar | bar in icon | arc | disc. arc | donut | disc. donut | pie | gauge | sliding |
| calories | | | | | 1 | | | | | |
| distance | | | | | 1 | | | | | 1 |
| heart rate | | | | 1 | 2 | | 1 | | 5 | 1 |
| steps | 6 | 3 | | 8 | 3 | 2 | 4 | 2 | 10 | 5 |
| humidity | | | | | | 3 | | | | |
| phone bat. | | 1 | | 3 | 2 | | 1 | | 3 | 1 |
| watch bat. | 10 | 7 | 13 | 11 | 5 | 2 | 8 | 2 | 26 | 4 |
| Sum | 16 | 11 | 13 | 23 | 14 | 7 | 14 | 4 | 44 | 12 |

Table 1 shows how often each type of construction appeared and how it was represented. The bar chart was the only representation that was sometimes integrated into an icon (for example, to show the watch battery status). When multiple proportion representations were shown together in 57.7% of the cases the same type of representation was used; 7.7% of the watch faces used only text as a representation; 34.6% of the watch faces used different types of representations, for example, an arc chart and a gauge chart.

Based on our analysis, we chose to focus on one common rectilinear and one common circular construction: a bar chart (in Table 1 used 40 times in total) and a radial bar (donut) chart (in Table 1 used 21 times in total). Despite its frequency, we did not pick an arc-type representation because of its wide variety of designs and display sizes in practice and instead focused on designs for which results could more easily be applied on watch faces. Due to its frequency, we used the same type of representation for all complications shown together. In addition, we chose a continuous representation instead of discrete steps for its higher frequency of use.

### 3.3 Watch Face Design

In the design of our studies, we had to make several choices regarding how to draw the data presented on a watch face:

**Visualization representations to compare.** We chose a bar chart and radial bar (donut) chart as outlined in Section 3.2.

**Location of the representations.** While there are no restrictions to the design of watch faces, it is common for data to be displayed in round complications close to the center of the watch. We chose to use the top center position for the display of time in a digital format in all three studies (Fig. 1), and reserved the border of the display for an analog time distractor in Studies 2 and 3. We used the space in the left-middle, bottom-center, and right-middle of the watch, which are common locations for the display of complications.

**Number of representations.** Our work is motivated by realistic watch faces, on which people commonly display 2–5 data items (in addition to time) simultaneously [31]. Given the regions of the watch face we had reserved for our data display, we chose to use three representations as a middle ground to not overcrowd the display.

The time shown was randomly generated and used the European format from 00 to 24 hours for the digital time display. We showed the three `Bar` and `Radial` charts in a light green color in Studies 1 and 2, and pink for Study 3 to increase contrast when the watch was tilted. The background of the complication was dark gray and the rest of the bar was shown in a lighter gray, ensuring enough contrast. In addition to proportion representations, we also showed an absolute number and a unit on the complication. We chose three common categories: calories burned, step count, and minutes of activity. We set the goal for each category ourselves based on common guidelines for active minutes (45 min), calories burned (2,100 kcal), and step count (12,000 steps). We generated the bar with a width of 89 px and a height of 12 px, and the radial bar with a width of 3.5 px and a radius of 45 px, which gave us roughly the same amount of pixels for the radial bar as used for the bar chart.

### 3.4 Method

We used a 1 up / 3 down *transformed and weighted staircase* procedure [26, 32], a common technique to investigate the perception of visual stimuli in psychophysics. Stimuli consisted of three complications each showing a different data value: either one, two, or three of the complications showed a value above 66% ($> \frac{2}{3}$). Participants had to determine how many complications showed more than 66%. Each stimulus was shown for a pre-defined time (the *stimulus exposure duration*). After each stimulus exposure, four intervening images, to account for after images, were shown [26]. We counterbalanced the order of representations using a reduced Latin square design. We recruited participants to take part in only one of the three studies.

**Measures.** We determined the *stimulus exposure duration* using a transformed and weighted up/down staircase method. We increased the exposure duration by 100 ms ($\Delta^+$) after one incorrect response and decreased it by 200 ms ($\Delta^-$) after three correct responses. To reach the true threshold faster, we decreased the duration after each correct answer by 200 ms before participants made their first mistake. This procedure generally leads to a threshold with ~69% correct responses [32, p.124]. We used two termination criteria: 25 reversals or 180 stimuli. However, in all three studies, none of the participants terminated after 180 stimuli, all terminated after 25 reversals.

We calculated the *time threshold* for each condition by averaging the stimulus exposure time of the *reversal* points in the staircase: trials in which participants oscillate between decrease-increase and increase-decrease. For each condition, we averaged the mean time of all reversal points after the second reversal. We also calculated the *mean error rate* to inspect if we had reached the ~69% correct responses [23]. For each condition, we computed the error rate as the number of incorrect answers over all answers given.

**Stimuli Generation.** We generated random data so that each answer (one, two, or three) was equally likely. The random data was generated to ensure a minimum length and to make values above 66% easily detectable: values above 66% were between 71 and 100 px long on the watch and the values below 66% were between 33 and 63 px long. We counterbalanced the position of the target complication(s) using the following seven different positions: for correct answer = 1: ●○○, ○●○, ○○●; correct answer = 2: ●○●, ●●○, ○●●; correct answer = 3: ●●●. For each correct answer, we prepared 60 different stimuli resulting in 180 stimuli in total.

### 3.5 Procedure and Apparatus

Participants first signed a consent form and filled out a background questionnaire. They then read a printed description of the study, the different conditions, and the procedure for one trial. For each condition, participants first performed 10 training trials and continued with the condition until one of the two termination criteria was reached. We set the starting time for each staircase to 1,000 ms. A trial consisted of the stimulus being shown based on the set stimulus exposure duration, then showing of four intervening images each for 25 ms, then answer input by the participant, and then display of the correctness of the answer. After each condition, we asked participants if they had a certain strategy to perform the task. After participants finished all conditions, we asked them to rank all representations based on aesthetics, efficiency, and their confidence in completing the task. We conducted the study in a lab with artificial lighting and no direct sunlight to avoid different lighting conditions.

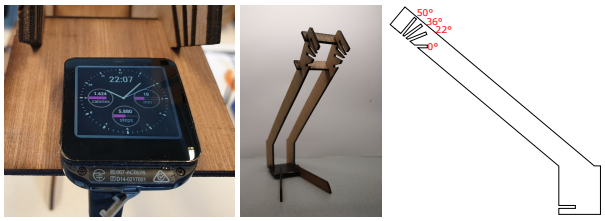



Figure 2: Left: The Sony Smartwatch 3 used in all three studies. Middle: Our custom-made study stand. Right: The angles we used in the study represented on a sketch of our study stand.

We used a Sony SmartWatch 3 (Fig. 2 Left) with Android Wear 2.8.0 as its operating system. The smartwatch has a screen resolution of 320 px × 320 px (= a pixel size of 0.089 mm) and a viewable screen area of 28.73 mm × 28.73 mm. As with previous studies [10, 11], we attached the smartwatch at an angle of 50° to a self-designed adjustable stand (Fig. 2 Middle). We adjusted the stand for each participant so that the smartwatch was at a height of 110 cm from the floor and at a viewing distance of 28 cm from the seated participant. We allowed participants to adjust their sitting position during the study but did not re-adjust the position of the stand. We placed a keyboard directly in front of participants, which they used to input answers using the arrow keys: ← (1 complication above 66%), ↓ (2 complications above 66%), → (3 complications above 66%).

Our study software recorded all key presses in a log file, calculated the stimulus exposure duration for each trial, and communicated with the smartwatch over TCP. If the termination criterion was reached, it started the next condition. The smartwatch was connected via a WiFi hotspot as a client and showed the stimuli and intervening images, as well as simple instructions and feedback.

### 3.6 Data Analysis and Interpretation

In all three studies, we analyzed the time threshold, accuracy, and participants' rankings of techniques. We report sample means of thresholds and 95% confidence intervals (CIs), which means we are 95% confident that this interval includes the population mean. We constructed all CIs using BCa bootstrapping (10,000 bootstrap iterations). We also report the CIs of mean differences to compare different conditions. The CIs of mean differences were adjusted for multiple comparisons with Bonferroni correction [29]. We analyzed the CIs using estimation techniques, which means that we interpret them as providing different strengths of evidence about the population mean, as recommended in the literature [6, 7, 17, 19, 20]. When reading a CI of mean differences, a non-overlap of the CIs with 0 is evidence of a difference, corresponding to statistically significant results in traditional $p$-value tests. Nevertheless, CIs allow for more subtle interpretations: the farther from 0 and the tighter the CI is, the stronger the evidence. Equivalent $p$-values can be obtained from CI results following Krzywinski and Altman [39].

In addition, we report subjective rankings of visualizations by participants. While these are not part of our main performance-focused research questions, the rankings provide additional elements that designers can consider when adding smartwatch complications.

Table 2: Details of participants for the three studies, all of whom participated in only one study. Number (#) of participants, gender (F: Female, M: Male), age, and number of participants who reported to own a wrist-worn device (WW). Familiarity was rated on a 5-point Likert scale (1: not familiar at all – 5: very familiar; MD: median, M: mean, SD: standard deviation). *One participant did not indicate their gender.

| Study | # | Gender | | Age | Occupant | Degree | WW | Familiarity | |
|---|---|---|---|---|---|---|---|---|---|
| | | F | M | (M / SD) | | | | Bar | Radial |
| **1** | 30 | 14 | 16 | 28.2/6.9 | 16 researchers, 5 engineers, 1 lecturer, 8 students | PhD (7), Master (16), Bachelor (7) | 5 | $MD = 5$, $M = 4.8$, $SD = 0.6$ | $MD = 4$, $M = 4.8$, $SD = 0.6$ |
| **2** | 30* | 6 | 23 | 25.1/7.0 | 3 researchers, 1 engineer, 26 students | Master (5), Bachelor (5), High School (20) | 5 | $MD = 5$, $M = 4.5$, $SD = 0.8$ | $MD = 4$, $M = 3.7$, $SD = 1.4$ |
| **3** | 14 | 5 | 9 | 26.7/6.4 | 7 researchers, 5 students, 2 social workers | Master (8), Bachelor (1), High School (5) | 3 | $MD = 5$, $M = 4.6$, $SD = 0.6$ | $MD = 4$, $M = 4.1$, $SD = 0.8$ |

## 4 Study 1: Digital Watch Face

Our first study investigated which of three representations (Bar, Radial, and Text70%) has the lowest time threshold using three complications and time represented in a digital format.

### 4.1 Study 1 Design Specifics

Here, we describe aspects unique to Study 1 compared to those reported in Section 3. In addition to Bar and Radial, we included a Text70% representation as a condition because it is commonly used to display information on watch faces [31]. In our survey, we found 63 of the proportion representations that also showed the proportion as text and an additional 52 items on the watch face that only showed the proportion as text. The stimuli used for this study are described in Section 3.4 and can be seen on the left of Fig. 1. We pre-registered Study 1 (`https://osf.io/xygj9`): we decided not to pre-register all three studies individually because we planned to have the same setup and analysis for all three studies.

### 4.2 Participants and Results

Table 2 summarizes background information about the participants we recruited for Study 1. Most (27) participants had human-computer interaction (13) and visualization (14) backgrounds, and three participants had a general computer science background. All participants had normal or corrected-to-normal vision and only one reported to be colorblind. Participants were not compensated, due to country regulations where the study was conducted, other than with a chocolate bar. Wrist-worn devices participants reported to own were made by Fitbit, Samsung, and Pebble.

The average completion time for the whole study, including reading the consent form, and filling in questionnaires, was 34 minutes. The time to finish all three conditions was on average 22 minutes.

**Thresholds.** The mean time thresholds for our representations were Bar = 318 ms [281 ms, 373 ms], Radial = 397 ms [337 ms, 473 ms], and Text70% = 503 ms [457 ms, 572 ms] (Table 4 Left; smaller values are better).

For the pair-wise differences of thresholds, we saw evidence of differences between all conditions. There is evidence that Radial was slower than Bar by 79 ms [17 ms, 160 ms], that in turn Text70% was slower than Radial by 106 ms [22 ms, 187 ms], and strong evidence that Text70% was slower than Bar by 185 ms [131 ms, 268 ms] (Table 4 Right). According to the obtained thresholds, the techniques rank: Bar < Radial < Text70%.

**Accuracy.** Following previous work [23], our study design aimed for ~69% correct responses (error of 31%). Mean error rates were lowest for Bar = 24% [23%, 25%], then Radial = 26% [25%, 28%], and Text70% = 31% [30%, 32%] (Table 5 Left).

Table 3: Ranking (RK) of the representations in Study 1 (left) & 2 (right): based on aesthetics (1st row), efficiency (2nd row), and confidence (3rd row). Ranking of the angles based on readability (4th row) for Study 3.

Ranking of Chart Aesthetics

| | Digital Watch Face | | | Digital+Analog Face | |
|---|---|---|---|---|---|
| RK | Bar | Radial | Text70% | Bar | Radial |
| 1 | 4 | 26 | 0 | 1 | 29 |
| 2* | 20 | 3 | 8 | 29 | 1 |
| 3* | 6 | 1 | 22 | | |

Ranking of Chart Efficiency

| | | | | | |
|---|---|---|---|---|---|
| 1 | 15 | 7 | 8 | 18 | 12 |
| 2 | 12 | 11 | 7 | 12 | 18 |
| 3 | 3 | 12 | 15 | | |

Confidence Ranking per Chart

| | | | | | |
|---|---|---|---|---|---|
| 1 | 16 | 7 | 7 | 17 | 13 |
| 2 | 10 | 13 | 7 | 13 | 17 |
| 3 | 4 | 10 | 16 | | |

*one participant ranked Bar and Text as equal

Watch Face Angle: Ranking of Angle Readability

| RK | 50° | 22° | 0° |
|---|---|---|---|
| 1 | 10 | 3 | 1 |
| 2 | 2 | 11 | 1 |
| 3 | 2 | 0 | 12 |

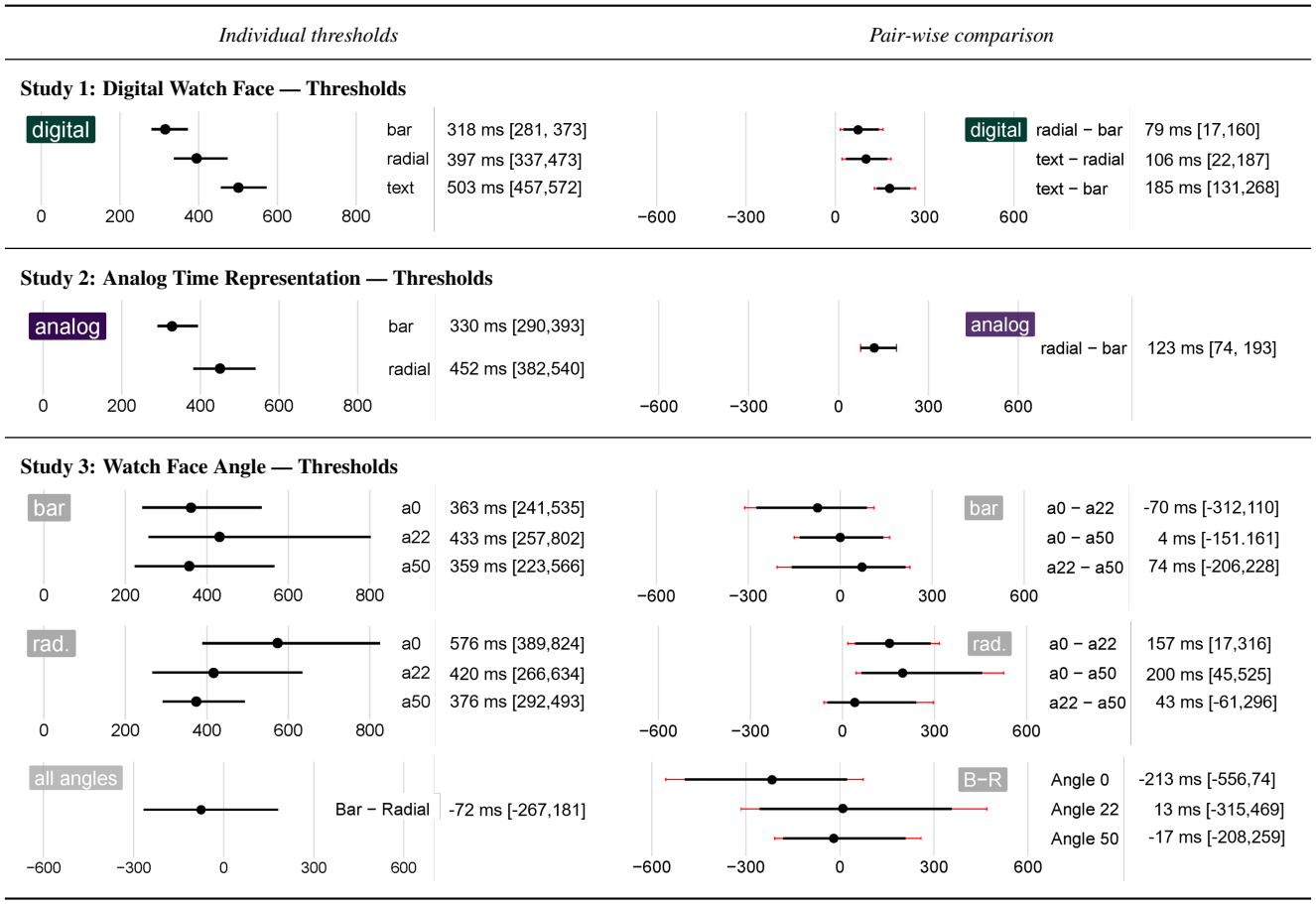
We have some evidence that based on the pair-wise error rate differences Radial was less accurate than Bar by 2% [1%, 4%], and strong evidence that Text70% was less accurate than both Radial by 4% [3%, 6%] and Bar by 7% [5%, 9%] (Table 5 Right). The ranking based on the errors is consistent with the threshold ranking: Bar < Radial < Text70%.

**Ranking.** Table 3 shows participants' subjective ranking for the three representations in terms of aesthetics, efficiency, and confidence in performing the task. Radial was considered the most aesthetically pleasing, followed by Bar and Text70% (Table 3 First row, Left). The rankings of perceived efficiency (Table 3 Second row, Left) and confidence (Table 3 Third row, Left) are consistent with those of the threshold and error measures, with Bar ranked first, followed by Radial, and Text70%.

## 5 Study 2: Adding an Analog Time Representation

Our survey showed that 28.3% of the watch faces used an analog time representation and 23.4% used both digital and analog (Section 3.2). In Study 2, we wanted to investigate if and how the addition of an analog time display affects reading time and error.

Table 4: Threshold analysis for all three studies. Left Row 1 and 2: Average thresholds in milliseconds for each *representation*. Right Row 1 and 2: Pair-wise comparisons for each *representation*. Left Row 3–5: Difference between independent means of the `Bar` and `Radial` thresholds across all *angles*. Right Row 3–5: Difference between independent means of the `Bar` and `Radial` thresholds for individual *angles*. Error bars represent 95% Bootstrap confidence intervals (CIs). In red are the CIs adjusted for three/two pairwise comparisons with Bonferroni correction.




### 5.1 Study 2 Design Specifics

We added an analog time representation to the digital display from Study 1. We dropped the text condition from Study 1 because it was the lowest ranked with the highest threshold and error rate.

We designed the analog watch face by placing ticks for each minute around the watch face with every 5 minutes as a thicker tick mark and a number for the hours next to it (Fig. 1 Middle). We placed the handles for the hour and minute so they would not overlap with the visualizations, ensuring that participants could read them. Therefore, we only used 12, 13, 22, and 23 as hours and minutes between 51 and 9. We refrained from using hours in the morning to avoid unrealistic data values such as 9,000 steps at 10 am.

The handles of an analog watch frequently cover a complication partially. We wanted to test this additional possible distraction from the task for added ecological validity. Therefore, we decided to place the handle for the seconds so that it would cover one of the complications. We used each of the values between 10 and 19, 25 and 34, as well as 40 and 49 to position the second handle.

### 5.2 Participants and Results

Table 2 summarizes participant detail in Study 2. Participants' backgrounds were: computer science (12) (of these 2 from visualization), computational linguistics (6), engineering (8), and other (4). All participants had normal or corrected-to-normal vision and only one participant reported to be colorblind. Participants received 10 € due to the study taking place in a country, which allowed compensation or a chocolate bar if they were employees of the university where we conducted the study. Wrist-worn devices participants reported to own were made by Apple, Sony, and Miband. The average time for finishing both conditions was 23 minutes.

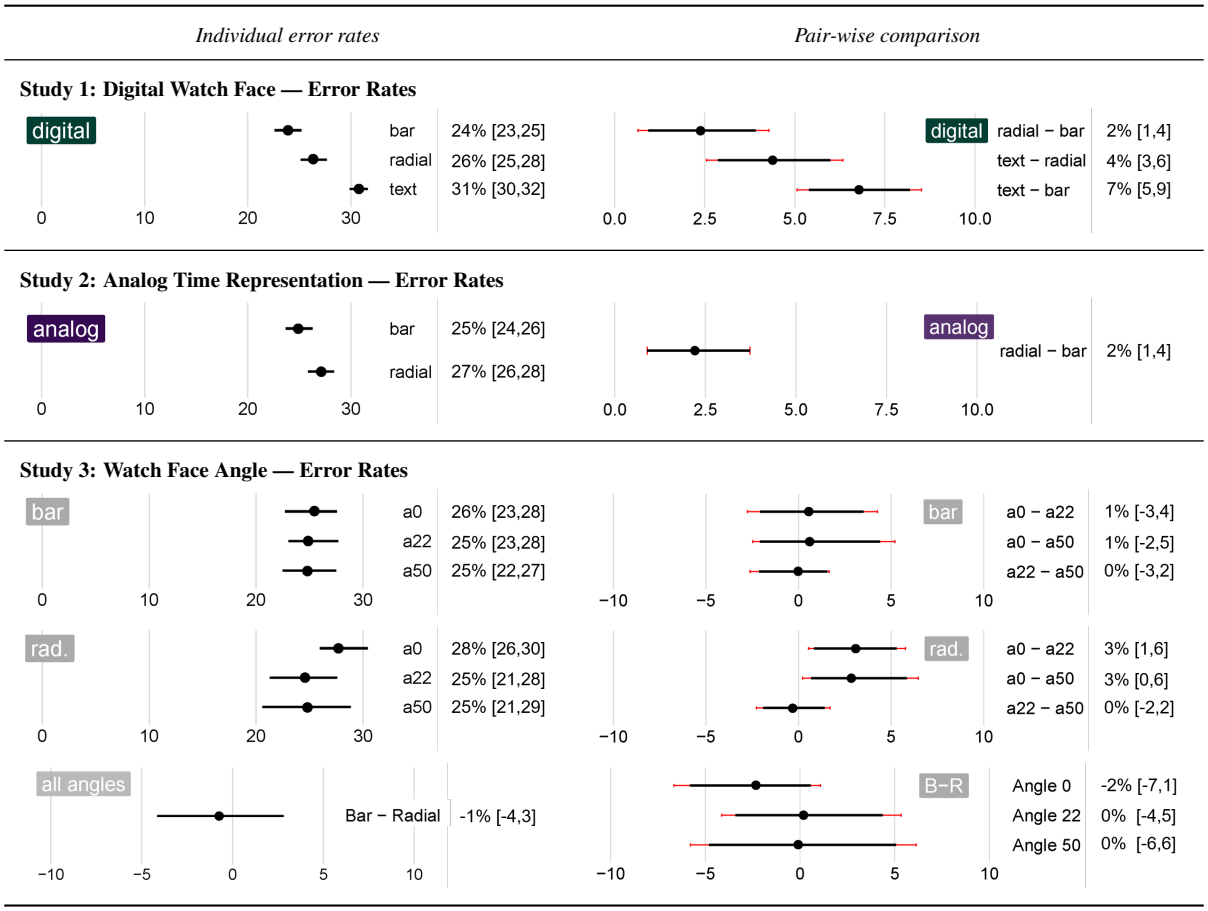
**Thresholds.** The time thresholds for the two representations were `Bar` = 330 ms [290 ms, 393 ms] and `Radial` = 452 ms [382 ms, 540 ms] (Table 4 Left). For pair-wise differences of thresholds, there is evidence that `Radial` was slower than `Bar` by 123 ms [74 ms, 193 ms] (Table 4 Right).

Compared to Study 1, the mean thresholds increased only slightly for both `Bar` (+12 ms) and `Radial` (+55 ms), indicating that the analog time representation did not slow down participants.

**Accuracy.** The mean error for the two representations were `Bar` = 25% [24%, 26%] and `Radial` = 27% [26%, 28%] (Table 5 Left). For pair-wise differences there is some evidence that `Bar` had a lower error than `Radial` by 2% [1%, 4%] (Table 5 Right). In comparison to Study 1, the error increased only slightly for `Bar` (+1%) and remained the same for `Radial`, indicating that the analog time representation did not affect error.

**Ranking.** Table 3 shows the average rankings for the different conditions in terms of aesthetics, efficiency, and confidence in performing the task. Overall, `Radial` was seen as more aesthetically pleasing (Table 3 First row, Right). However, for efficiency and confidence, the `Bar` was ranked higher (Table 3 Second & Third row, Right), results that are consistent with performance measures and Study 1.

Table 5: Error analysis of data from all three studies. Left Row 1 and 2: Average error rate for each *representation*. Right Row 1 and 2: Pair-wise comparisons for each *representation*. Left Row 3–5: Difference between independent means of the `Bar` and `Radial` mean errors across all *angles*. Right Row 3–5: Difference between independent means of the `Bar` and `Radial` mean errors for individual *angles*. Error bars represent 95% Bootstrap confidence intervals (CIs). In red are the CIs adjusted for three/two pairwise comparisons with Bonferroni correction.




## 6 Study 3: Watch Face Angle

Blascheck et al. [10, 11] investigated the average angles at which people held their watch when reading information. They reported an average angle of a worn watch of 50° (*SD* = 14°) with the captured angles spanning a range of 36°–64°. Given the wide angle difference, we wanted to investigate what effect the viewing angle would have on the thresholds and accuracy of proportion judgments in a less-controlled scenario with increased external validity.

### 6.1 Study 3 Design Specifics

Our goal was to find a minimum time threshold for three different viewing angles. Unlike Studies 1 and 2, we set up a between-subject experiment and used a chin rest to ensure that all participants had the same viewing angle. We used 36° (1 *SD* from the average angle) for training. We included the 50° viewing angle that we used in Studies 1 and 2 to compare trends from the two previous studies. We chose 0° as an extreme case, in which the smartwatch is oriented parallel to the table. We also included the 22° angle, which is 2 *SD* from the average angle and roughly in the middle between 0° and 50°. Fig. 2 Right shows all four angles.

We used the time and stimulus design as in Study 2 (Section 5.1) but had to adjust the complication design. At an angle of 0°, the smartwatch screen exhibited a considerable loss in brightness and contrast ratio that required us to change the color of the visualizations to a bright magenta, remove the background of the complications, and desaturate the complication border (Fig. 1 Right).

### 6.2 Participants

Table 2 shows details about the participants. Due to the Covid-19 pandemic, we had to stop the study after 14 participants to ensure the safety and health of participants and experimenter. Even though 14 participants are less than the 36 we initially planned, our results highlight potential trends from this smaller sample. There is no magic number for participants required in a study [4] and visualization studies often have small numbers of participants with relevant results [8, 13]. When it comes to statistical evidence, CIs with just two participants can still provide evidence of differences [20].

Participants' backgrounds were: visualization (4), human-computer interaction (2), general computer science (2), engineering (2), social worker (2), and other (2). All had normal or corrected-to-normal vision without a vision deficit. Participants received 10 € or a chocolate bar (if employed by the university where we conducted the study), as per the local regulations. Wrist-worn devices participants reported to own were made by Garmin and Apple.

### 6.3 Results

To compare angles within each visualization, we used bootstrap CI calculations (adjusted for multiple comparisons). To compare the `Bar` and `Radial` stimuli conditions (between-subjects) we used bootstrap CI calculations for two independent samples.

The completion time for finishing all four conditions (including the training condition) was on average 29 minutes.

**Thresholds.** First, we investigate the effect of angle for each representation: for Bar the threshold means were smallest for 50° = 359 ms [223 ms, 566 ms] followed by 0° = 363 ms [241 ms, 535 ms], and then 22° = 433 ms [257 ms, 802 ms] (Table 4 Left). From the CIs of our sample and the pair-wise comparisons of the viewing angles, we did not find evidence of true differences between these means (Table 4 Right). Given our sample size, we cannot exclude that trends could emerge with more participants; however, considering our results so far, we do not suspect large effects. The detailed CI values can be found in Table 4.

For Radial, some trends emerge. Mean threshold times were smallest for 50° = 376 ms [292 ms, 493 ms], then 22° = 420 ms [266 ms, 634 ms], and finally 0° = 576 ms [389 ms, 824 ms] (Table 4 Left). Pair-wise mean comparisons provide evidence that the viewing angle 50°, and possibly 22°, are likely faster than the extreme viewing angle 0° (Table 4 Right).

These early results indicate that the thresholds for Bar seem not to be greatly affected by viewing angles. Radial, however, is affected by extreme viewing angles (0° in our case), but we were unable to find differences across other angles. If other differences exist for the remaining angles, they will likely be very small. This indicates that results from studies that use fixed angles, such as our Studies 1 & 2, as well as those from previous work on smartwatch glanceability [10], probably hold for a wider range of viewing angles.

The between-subject difference of the independent means of Bar and Radial across all angles was -72 ms [-267 ms, 181 ms], and we do not have evidence of a difference between the two representations (Table 4 Left). For the individual angles, there was also no evidence of a difference between Bar and Radial (Table 4 Right). However, for 0° there may be a trend that Bar has a slightly lower threshold than Radial (-213 ms [-556 ms, 74 ms]), in other words: Bar may be more robust to extreme viewing angles.

**Accuracy.** Considering accuracy for Bar, mean error rates were fairly similar for all angles, with 50° = 25% [22%, 27%] and 22° = 25% [23%, 28%] having the lowest error rates, followed by 0° = 26% [23%, 28%] (Table 5 Left). We do not have strong evidence of differences across angles (Table 5 Right).

Considering accuracy for Radial, mean error rates were lowest for 50° = 25% [21%, 29%] and 22° = 25% [21%, 28%], and with 0° = 28% [26%, 30%] being the most error prone (Table 5 Left). Looking at pair-wise comparisons, there may be a trend that the extreme viewing angle 0° is more error prone than 22° and 50° (Table 5 Right).

Overall, mean error rate differences were small across the board (between 1-3%). For Radial, our primary results indicate that the extreme viewing angle 0° may be more error prone than all others. Consistent with the threshold results, there is neither evidence of a difference in error rate between the two visualizations overall (Table 5 Left), nor when considering individual angles (Table 5 Right).

**Ranking.** Participants also ranked the three angles from best (1) to worst (3) based on their ability to read them (Table 3 Fourth row). 50° was ranked first the most (10 votes), 22° was ranked second the most (11), and 0° was ranked third the most (12).

## 7 Discussion

In this section, we reflect on our research questions, results, and their design implications. We end with listing limitations of our studies and possibilities for future work.

### 7.1 Summary & Design Implications (DIs)

**(RQ1) What are common watch face representations to represent proportions?** Our pre-study showed that two main rules of construction were used for smartwatch proportion representations: rectilinear and circular. Bar charts were the most common rectilinear and arcs, gauges, donuts, and pies for circular construction. While arcs were most common they showed a huge variety of designs and sizes, and were often used against the rim of a round watch face, while radial bars were more standardized in design. Some of the representations took the form of unit-based visualizations, broken into smaller chunks, but the majority were continuous representations. When multiple types of data were shown on a watch face most commonly the same representation type was used, for example, bar charts to represent steps and calories.

→ ***DI 1:*** *Our survey showed what was common on watch faces but also pointed out avenues of future research, which we discuss in the next section. Whether or not a designer wants to follow what is common is a personal choice that involves considering how used people are to seeing certain representations or their combinations on the same watch face—in contrast to possible novelty effects or more engaging designs that have not yet been frequently used.*

**(RQ2) Which part-to-whole proportion representation has the lowest time threshold and highest accuracy for multiple smartwatch complications?** In Study 1, we found strong evidence that visual representations outperformed Text70% proportion representations. Bar and Radial representations were faster with less error and ranked higher by participants for aesthetics, efficiency and confidence. This reflects previous work showing that visualization can outperform text to display proportions (e.g., Spence and Lewandowsky [48]). The data we gathered also highlights differences between Bar and Radial representations, albeit both the evidence and the effects are weaker: Bar performed slightly better than Radial in both time and accuracy of reading. Previous studies on smartwatches [10] found a slight advantage for donut charts in data comparison tasks, so our results are different, although in that previous work, the differences were also marginal. The ranking data reflects this marginal difference too: Bar was perceived as more efficient and giving more confidence to participants; however, Radial was preferred for being more aesthetically pleasing. Other visualization work that did not focus on smartwatches found pie and donut charts to perform similarly [38]. Should this work generalize to smartwatches, our work suggests that bar charts will perform equally or better than pie charts (not just donuts) for part-to-whole tasks, consistent with results from previous work (see Section 2.1). Our results are also similar to Blascheck et al.'s study [10] which showed that the studied task could be completed with bar and donut charts on average in <300 ms. Our results here had average thresholds of ~300 ms for both Bar and Radial.

→ ***DI 2:*** *We recommend not using pure text representations for multiple proportion estimations. However, for other types of tasks and data (e. g., real-time heart rate), future studies are needed.*

→ ***DI 3:*** *The performance of Bar and Radial charts was quite similar, with a slight advantage for Bar (~80 ms) at the cost of a decrease in perceived aesthetics. Given how small these differences were, designers can choose between a slightly more efficient chart and a more visually pleasing one.*

**(RQ3) How does the density and complexity of a watch face affect task performance?** Adding a display of time in an analog format increases the visual complexity of a watch face. A first step to increase the external validity of our work was to see if our results would hold if we were to add an analog time display. Our data suggests that the general trend of Bar slightly outperforming Radial also held on a watch face with an analog display of time. The trends observed in the first study also held for self-reported ranking: participants were more confident with Bar and felt more efficient, while Radial was rated most aesthetically pleasing. Putting the results of both studies side by side, the performance (accuracy and time) might practically only slightly deteriorate when an analog time display is added. The average differences we observed were small (<55 ms). The impact of other types of visual clutter (animation, more complications, color variation, or text) still

needs to be studied because they might have different effects on task performance (see the literature on visual clutter [22, 44]).

→ ***DI 4:*** *Analog displays of time are present on roughly one third of smartwatch faces [31] (https://osf.io/nwy2r/) and as such a significant number of people see watch faces with added visual complexity. Our results suggest that adding an analog time display had only a negligible effect on reading multiple proportion representations. In situations in which a small viewing time difference (<55 ms in our study) matters for the detection of goal completion progress, designers could automatically switch from an analog to a digital time display to simplify the watch face.*

**(RQ4) How does the viewing angle of a smartwatch affect task performance?** In practice, the viewing angle of a smartwatch is unlikely to be fixed. It seems that for Bar viewing angles do not have an effect on performance. For Radial, our results suggest slightly worse performances for 0° (parallel to the floor). With both representations, angles of 22° and 50° obtained relatively comparable results. In Studies 1 and 2, we chose an angle of 50° based on previous studies [10, 11]. Our preliminary results for Study 3 suggest that our results with 50° would hold with other viewing angles, including the range between 22° and 50°.

→ ***DI5:*** *Our findings indicate that only extreme viewing angles affect performance in Radial, while multiple Bar representations seem robust to viewing angle changes. Designers can assume that these visualizations can be read under different viewing angle conditions; however, if high accuracy is required they should choose Bar. We also note that the findings of past studies using fixed angles likely hold for larger angle ranges.*

**Summary:** Our studies found that text representations on smartwatch faces are less glanceable and accurate in comparison to Bar or Radial charts. However, which type of representation to choose can remain up to the wearer or designer, because performance differences are small. The complexity of the smartwatch face using both an analog and digital time display only has a small impact on performance. We also saw that only the extreme angle affected acted the readability of the watch face.

### 7.2 Limitations and Future Work

Our results should be considered as a lower bound estimate of the performances of a general population in a realistic usage scenario. This is a common yet inevitable limitation of controlled experiments. A realistic usage scenario for a smartwatch would entail a variety of viewing angles, a moving watch, different lighting and viewing conditions, and varying cognitive load on the wearer. All of these factors might affect performance negatively. We address this limitation by providing first results on viewing angles that show that prior results would probably hold for different viewing angles. While assessing in more detail how our results hold across more angles and distances requires further experimentation; our current study indicates that fixed and non-extreme angles do not seem to affect performances. We expect that the higher cognitive load in more demanding contexts in the real world (walking, talking, or running) will more seriously affect performance [9, 15, 52]. Testing such effects requires experimental protocols that reliably control the cognitive load of participants to emulate everyday situations. This remains a challenging, but exciting future research direction that can bring together researchers from visualization and cognitive sciences.

As is often the case with in-person visualization studies that need access to a specific setup, we relied mostly on students and employees from our university, who often have a high education background and may be more aware of novel devices than an average person in the population. Nevertheless, their qualifications (younger, higher education, tech-savvy) allow us to posit that our samples would perform better than a representative sample of the global population. Consequently, we should treat the values that we have found as a lower estimate of the true population mean. In other words, with a more diverse sample (e. g. older adults), the exact values would be higher but the trends observed will likely remain, making our results and design recommendations still valid, as none of them are based on actual values but rather on the effects observed.

On a related note, a third limitation lies in the relatively small sample size of our last study, due to its early termination. We have mitigated this limitation by acknowledging in our results that our evidence is weaker due to this smaller sample size, and adapting our language to reflect the greater uncertainty insisting our results are preliminary. Nevertheless, we note that our results showing differences across visualizations in this third study remain valuable, because there is no ideal sample size for a study [3, 4], statistical CI tests can show trends with very few participants [20], and that relatively small sample sizes such as ours are common in many visualization and human-computer interaction studies [8, 13, 33].

Other real-use factors provide starting points for future research. These include the number, nature, and placement of the different visualizations and complications. In a recent smartwatch survey [31], most wearers reported having 2–5 data items on their display. Our choice of three complications falls within this common-use range, nevertheless future work should consider some of the extreme numbers and their effect on glanceability. Similarly, the survey revealed that watch faces sometimes do combine several different visual encodings together. Our studies only considered complications with the same encoding, and mixing encodings can affect performance. Studying how combining visualizations with different encodings may affect glanceability remains an open question for smartwatches, but also for other visualization displays. Finally, in particular arcs and gauges revealed that complications vary in both size and placement (as can also be seen in the survey teaser of Islam et al. [31]). This diversity of forms, sizes, and placements opens a large number of potential variations to study in the future.


## Acknowledgments

This work is funded by ANR grant ANR-18-CE92-0059-01 and DFG grant ER 272 14-1. Tanja Blascheck is funded by the European Social Fund and the Ministry of Science, Research, and Arts Baden-Württemberg. Lonni Besançon is funded by the Knut and Alice Wallenberg Foundation (grant KAW 2019.0024).


## References


[1] F. Amini, K. Hasan, A. Bunt, and P. Irani. Data representations for in-situ exploration of health and fitness data. In *Proc. of PervasiveHealth*, pp. 163–172. ACM, 2017. doi: 10.1145/3154862.3154879

[2] R. Aravind, T. Blascheck, and P. Isenberg. A survey on sleep visualizations for fitness trackers. In *Proc. of EuroVis – Posters*, pp. 85–87, 2019.

[3] A. Bacchelli and M. Beller. Double-blind review in software engineering venues: The community's perspective. In *IEEE/ACM Int. Conf. on Software Engineering Companion*, pp. 385–396. IEEE/ACM, 2017. doi: 10.1109/ICSE-C.2017.49

[4] P. Bacchetti. Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine*, 8(1), 2010. doi: 10.1186/1741-7015-8-17

[5] J. Bertin. *Semiology of Graphics: Diagrams Networks Maps*. Esri Press, 2011.

[6] L. Besançon and P. Dragicevic. The significant difference between p-values and confidence intervals. In *Proc. of the Conf. on l'Interaction Homme-Machine*, pp. 53–62, 2017.

[7] L. Besançon and P. Dragicevic. The continued prevalence of dichotomous inferences at CHI. In *Extended Abstracts of the Conf. on Human Factors in Computing Systems*, pp. 1–11. ACM, 2019.

[8] L. Besançon, A. Ynnerman, D. F. Keefe, L. Yu, and T. Isenberg. The state of the art of spatial interfaces for 3D visualization. *CGF*, 40(1):293–326, 2021. doi: 10.1111/cgf.14189

[9] M. Blakely, K. Wilson, P. Russell, and W. Helton. The impact of cognitive load on volitional running. In *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, pp. 1179–1183, 2016. doi: 10.1177/1541931213601276

[10] T. Blascheck, L. Besançon, A. Bezerianos, B. Lee, and P. Isenberg. Glanceable visualization: Studies of data comparison performance on smartwatches. *IEEE TVCG*, 25(1):630–640, 2019. doi: 10.1109/TVCG.2018.2865142

[11] 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 *Proc. of the Workshop on Data Visualization on Mobile Devices held at ACM CHI*, 2018.

[12] X. Cai, K. Efstathiou, X. Xie, Y. Wu, Y. Shi, and L. Yu. A study of the effect of doughnut chart parameters on proportion estimation accuracy. *CGF*, 37(6):300–312, 2018. doi: 10.1111/cgf.13325

[13] K. Caine. Local standards for sample size at CHI. In *Proc. of the Conf. on Human Factors in Computing Systems*, pp. 981–992. ACM, 2016. doi: 10.1145/2858036.2858498

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

[15] H. Cho, N. Romine, F. Barbieri, and S. Rietdyk. Gaze diversion affects cognitive and motor performance in young adults when stepping over obstacles. *Gait & Posture*, 73:273–278, 2019. doi: 10.1016/j.gaitpost.2019.07.380

[16] E. K. Choe, B. Lee, and m.c. schraefel. Characterizing visualization insights from quantified selfers' personal data presentations. *IEEE CG&A*, 35(4):28–37, 2015. doi: 10.1109/MCG.2015.51

[17] A. Cockburn, P. Dragicevic, L. Besançon, and C. Gutwin. Threats of a replication crisis in empirical computer science. *Comm. of the ACM*, 63(8):70–79, 2020. doi: 10.1145/3360311

[18] F. Croxton and R. Stryker. Bar charts versus circle diagrams. *J. Am. Stat. Assoc.*, 22(160):473–482, 1927. doi: 10.2307/2276829

[19] G. Cumming. *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge, 2013.

[20] P. Dragicevic. Fair statistical communication in HCI. In J. Robertson and M. Kaptein, eds., *Modern Statistical Methods for HCI*, chap. 13, pp. 291–330. Springer, 2016. doi: 10.1007/978-3-319-26633-6_13

[21] W. Eells. The relative merits of circles and bars for representing component parts. *J. Am. Stat. Assoc.*, 21(154):119–132, 1926. doi: 10.1080/01621459.1926.10502165

[22] G. Ellis and A. Dix. A taxonomy of clutter reduction for information visualisation. *IEEE TVCG*, 13(6):1216–1223, 2007. doi: 10.1109/TVCG.2007.70535

[23] 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

[24] R. Gouveia, E. Karapanos, and M. Hassenzahl. How do we engage with activity trackers? a longitudinal study of habito. In *Proc. of UbiComp*, pp. 1305–1316. ACM, 2015. doi: 10.1145/2750858.2804290

[25] R. Gouveia, F. Pereira, E. Karapanos, S. A. Munson, and M. Hassenzahl. Exploring the design space of glanceable feedback for physical activity trackers. In *Proc. of UbiComp*, pp. 144–155. ACM, 2016. doi: 10.1145/2971648.2971754

[26] M. Greene and A. Oliva. The briefest of glances: The time course of natural scene understanding. *Psychological Science*, 20(4):464–472, 2009. doi: 10.1111/j.1467-9280.2009.02316.x

[27] C. G. Healey and A. P. Sawant. On the limits of resolution and visual angle in visualization. *ACM Transactions on Applied Perception*, 9(4), Oct. 2012. doi: 10.1145/2355598.2355603

[28] 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 *Proc. of the Conf. on Human Factors in Computing Systems*, pp. 1303–1312. ACM, 2009. doi: 10.1145/1518701.1518897

[29] J. Higgins. *Introduction to Modern Nonparametric Statistics*. Thomson Learning, 2004.

[30] J. Hollands and I. Spence. Judging proportion with graphs: The summation model. *Applied Cognitive Psychology*, 12(2):173–190, 1998. doi: 10.1002/(SICI)1099-0720(199804)12:2<173::AID-ACP499>3.0.CO;2-K

[31] A. Islam, A. Bezerianos, B. Lee, T. Blascheck, and P. Isenberg. Visualizing information on watch faces: A survey with smartwatch users. In *Proc. of IEEE VIS – Short Papers*, pp. 156–160, 2020. doi: 10.1109/VIS47514.2020.00038

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

[33] L. Koeman. How many participants do researchers recruit? a look at 678 ux/hci studies. Website. Last visited: January 2023, https://lisakoeman.nl/blog/how-many-participants-do-researchers-recruit-a-look-at-678-ux-hci-studies.

[34] R. Kosara. Circular part-to-whole charts using the area visual cue. In *Proc. of EuroVis – Short Papers*, pp. 13–17, 2019.

[35] R. Kosara. Evidence for area as the primary visual cue in pie charts. In *Proc. of IEEE VIS – Short Papers*, pp. 101–105, 2019. doi: 10.1109/VISUAL.2019.8933547

[36] R. Kosara. The impact of distribution and chart type on part-to-whole comparisons. In *Proc. of EuroVis – Short Papers*, pp. 7–11, 2019.

[37] R. Kosara and D. Skau. Judgment error in pie chart variations. In *Proc. of EuroVis – Short papers*, pp. 91–95, 2016.

[38] R. Kosara and C. Ziemkiewicz. Do mechanical turks dream of square pie charts? In *Proc. of BELIV: Beyond time and errors: Novel evaluation methods for information visualization*, pp. 63–70, 2010.

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

[40] Little Labs, Inc. Facer - thousands of free watch faces for apple watch, samsung gear s3, huawei watch, and more. `https://www.facer.io/`. Last visited: January 2023.

[41] D. McMillan, B. Brown, A. Lampinen, M. McGregor, E. Hoggan, and S. Pizza. Situating wearables: Smartwatch use in context. In *Proc. of the Conf. on Human Factors in Computing Systems*, pp. 3582–3594, 2017. doi: 10.1145/3025453.3025993

[42] A. Neshati, Y. Sakamoto, L. Leboe-McGowan, J. Leboe-McGowan, M. Serrano, and P. Irani. G-Sparks: Glanceable sparklines on smartwatches. In *Proc. of the Graphics Interface Conf.*, 2019. doi: 10.20380/GI2019.23

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

[44] R. Rosenholtz, Y. Li, J. Mansfield, and Z. Jin. Feature congestion: A measure of display clutter. In *Proc. of the Conf. on Human Factors in Computing Systems*, pp. 761–770. ACM, 2005. doi: 10.1145/1054972.1055078

[45] D. Simkin and R. Hastie. An information-processing analysis of graph perception. *J. Am. Stat. Assoc.*, 82(398):454–465, 1987. doi: 10.1080/01621459.1987.10478448

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

[47] I. Spence and P. Krizel. Children's perception of proportion in graphs. *Child Development*, 65(4):1193–1213, 1994. doi: 10.2307/1131314

[48] I. Spence and S. Lewandowsky. Displaying proportions and percentages. *Applied Cognitive Psychology*, 5(1):61–77, 1991. doi: 10.1002/acp.2350050106

[49] A. C. Suciu and J. E. Larsen. Active self-tracking and visualization of subjective experience using vas and time spirals on a smartwatch. In *Proc. of the Workshop on Data Visualization on Mobile Devices held at ACM CHI*, 2018.

[50] 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 *Proc. of the Conf. on Human Factors in Computing Systems*, pp. 3569–3581. ACM, 2017. doi: 10.1145/3025453.3025817

[51] Wikimedia Foundation, Inc. Complication (horology). `https://en.wikipedia.org/wiki/Complication_(horology)`. Last visited: January, 2023.

[52] G. Yogev-Seligmann, Y. Rotem-Galili, A. Mirelman, R. Dickstein, N. Giladi, and J. M. Hausdorff. How does explicit prioritization alter walking during dual-task performance? effects of age and sex on gait speed and variability. *Physical Therapy*, 90(2):177–186, 2010. doi: 10.2522/ptj.20090043

[53] C. Ziemkiewicz and R. Kosara. Implied dynamics in information visualization. In *Proc. Adv. Vis. Interf.*, pp. 215–222. ACM, 2010. doi: 10.1145/1842993.1843031