Visualizing Ranges over Time on Mobile Phones: A Task-Based Crowdsourced Evaluation

Matthew Brehmer, Bongshin Lee, Petra Isenberg, Eun Kyoung Choe

Visualizing Ranges over Time on Mobile Phones: A Task-Based Crowdsourced Evaluation

Matthew Brehmer, Bongshin Lee, Petra Isenberg, and Eun Kyoung Choe

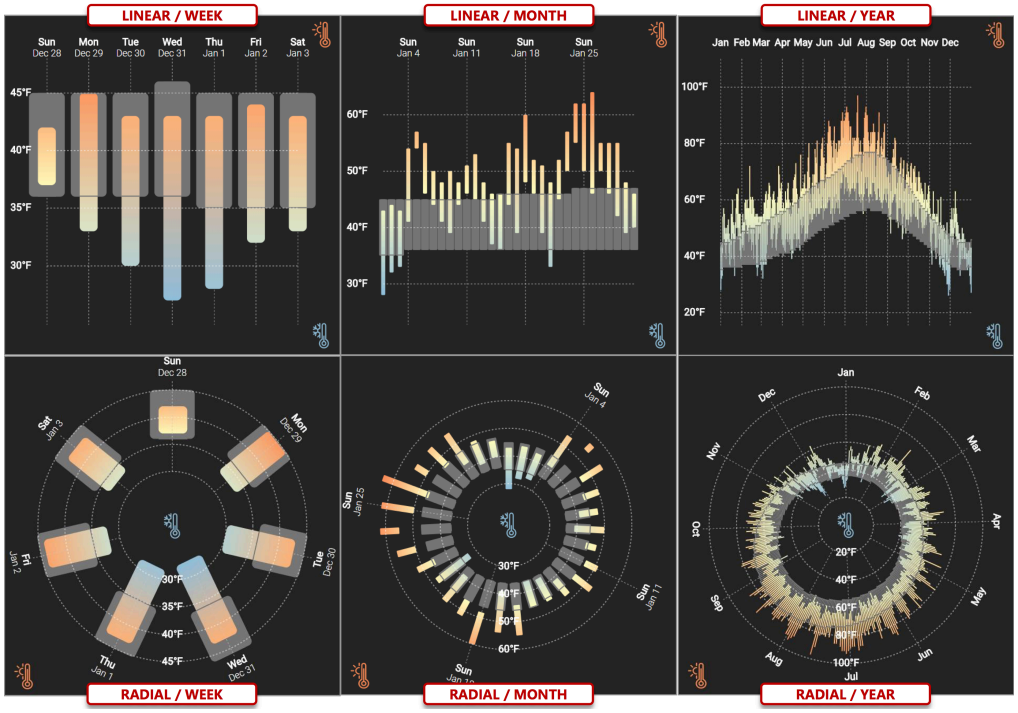


Fig. 1. *Linear* and *Radial* temperature range charts designed for mobile phone displays, representative of the stimuli used in our crowdsourced experiment. The colored bars encode observed temperature ranges and are superimposed on gray bars encoding average temperature ranges. Corresponding *Week*, *Month*, and *Year* charts display the same data (Seattle temperatures in 2015).

Abstract—In the first crowdsourced visualization experiment conducted exclusively on mobile phones, we compare approaches to visualizing ranges over time on small displays. People routinely consume such data via a mobile phone, from temperatures in weather forecasting apps to sleep and blood pressure readings in personal health apps. However, we lack guidance on how to effectively visualize ranges on small displays in the context of different value retrieval and comparison tasks, or with respect to different data characteristics such as periodicity, seasonality, or the cardinality of ranges. Central to our experiment is a comparison between two ways to lay out ranges: a more conventional linear layout strikes a balance between quantitative and chronological scale resolution, while a less conventional radial layout emphasizes the cyclicality of time and may prioritize discrimination between values at its periphery. With results from 87 crowd workers, we found that while participants completed tasks more quickly with linear layouts than with radial ones, there were few differences in terms of error rate between layout conditions. We also found that participants performed similarly with both layouts in tasks that involved comparing superimposed observed and average ranges.

Index Terms—Evaluation, graphical perception, mobile phones, range visualization, crowdsourcing.

1 Introduction

With the proliferation of smartphones and mobile apps, it has become commonplace to consume quantitative information on a mobile device. People consult apps and websites on their phone to monitor and compare quantities pertaining to weather, finance, personal health, and countless other interests. Designers often represent these quantities visually to facilitate skimming, comparison, and the identification of trends and outliers. While extensive research has been conducted to understand how people perceive quantities using charts on desktop displays, mobile phones have less physical screen real-estate with unique aspect ratios, and are viewed in various viewing conditions. So far, we have little understanding of how a mobile phone context influences the effectiveness of visualization. In this paper, we study how people experience visual representations of quantitative information when constrained by the size and aspect ratios of small displays. Specifically, we experimentally investigate charts that display ranges over time, a type of information appearing often on mobile phones.

- *Matthew Brehmer and Bongshin Lee are with Microsoft Research. E-mail: {mabrehme, bongshin}@microsoft.com.*
- *Petra Isenberg is with Inria. E-mail: petra.isenberg@inria.fr.*
- *Eun Kyoung Choe is with The University of Maryland, College Park. E-mail: choe@umd.edu.*

Manuscript received 31 Mar. 2018; accepted 11 Jul. 2018. Date of Publication 1 Aug. 2018; date of current version 1 Aug. 2018. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

A range is a pair of quantitative values, such as the low and high temperature of a single day, or the minimum and maximum selling price of a stock in a financial quarter. To present multiple ranges where each range is associated with a unit of time, a common approach is to visually encode these ranges as bars or pairs of points, positioned relative to one quantitative axis and one chronological axis (e.g., see the top row of Figure 1, where the former axis is vertical while the latter is horizontal). Among many factors to consider when visualizing ranges over time, we address four of them in a crowdsourced experiment conducted with 87 crowd workers using their own mobile phones:

- *Layout:* A more conventional *Linear* layout having a balanced quantitative and chronological scale resolution versus a less conventional *Radial* layout that emphasizes the cyclicality of time and may prioritize discrimination between values at its periphery.
- *Data source:* Two sources of range data (a city's daily high & low temperatures, and a person's sleep schedule of bedtimes & waking times), both being representative of data consumed on a phone.
- *Granularity:* Three granularities of time (*Week*, *Month*, and *Year*) that correspond with an increasing cardinality of ranges and thus a higher density of daily range marks (7, 28−31, and 365−366).
- *Task:* A representative set of five value retrieval and comparison tasks that people are likely to perform with ranges over time.

We contribute findings from our experiment, which reveal performance differences between *Linear* and *Radial* layouts of ranges over time in the context of data characteristics, granularities of time, and different tasks. Overall, while participants completed tasks faster with *Linear* layouts than with *Radial* ones, both layouts generally incurred a similar number of errors. Participants also universally preferred *Linear* layouts. Given these findings, we discuss design implications for mobile use cases involving the visualization of ranges. From a methodological standpoint, we build upon previous work involving the crowdsourcing of graphical perception [32] to offer the first crowdsourced visualization experiment conducted exclusively on mobile phones.

2 Background and Related Work

We were inspired by the rise of mobile apps incorporating visualization, as well as by alternative approaches for visualizing ranges, which include work by practitioners and work published in the research literature. Our work also draws from existing research on visualizing data on small displays and from previous visualization evaluation studies.

2.1 Visualization on Mobile Devices

In recent years, we have seen an increase in the number of mobile apps that feature visual representations of data. With the survey efforts of Ros et al. [50] and Sadowski [51] as well as a recent discussion on data-driven storytelling across different devices [43], the visualization community has begun to pay attention to the design space for visualizing data on mobile devices. However, more work is needed to evaluate design choices for mobile devices [10], to provide guidance that is particular to characteristics of the data and the context of use.

Mirroring the rise of visualization on mobile devices, the body of research literature proposing and evaluating visual encoding and interaction techniques for mobile devices has also emerged in recent years [53], including efforts to establish a research agenda in this area [42]. This body of work spans different datatypes, including continuous time-series data [18], hierarchical data [3], and spatiotemporal data [39], as well as various application areas, from healthcare [19] to navigation [36]. Others have considered the design of touch interactions for data visualization on tablet devices (e.g., [8, 27, 34]). While we do not focus on mobile interaction design nor tablet devices in this work, we empirically evaluate design choices for a datatype that has yet to be examined on mobile phones, that of ranges over time.

2.2 Visualizing Ranges

A range is a pair of values sharing a common quantitative domain. While a range can be encoded using two marks positioned along a one-dimensional scale, with one mark for the start of the range and the other for its end, it is more common to encode a range with a rectangular mark spanning part of the scale. When multiple ranges are placed adjacent to one another, a viewer can perform position comparisons along the shared scale. Unlike a bar chart, a "range chart" comprised of multiple rectangular marks would form what Fuchs et al. refer to as a profile without a common baseline [29], in that the rectangles need not be aligned to the start of the scale's domain (such as in the top row of Figure 1). In this sense, a range chart is visually similar to a simplified version of a boxplot [54], one that encodes the start and end of ranges rather than a full set of distribution statistics. While there may be other visual channels with which to encode a set of ranges, such as the "color stock chart" used in Albers et al.'s evaluation of aggregation techniques for time-series data [4], we focus primarily on position along a common scale, a visual channel that is known to be more effective for quantitative judgments relative to other channels [47].

In practice, range charts incorporating rectangular marks are fairly common, particularly in weather reporting. For instance, the print edition of the *New York Times* features a small multiple range chart encoding high and low temperatures for 22 major American cities for the five preceding days and the next five days, which are superimposed over the average and record high & low temperatures for those same days. Elsewhere, we have encountered range charts spanning seven months [11] up to an entire year [49]. Kekeritz's "Weather Radials" [37] presents 365 days of recorded and average temperature ranges for 35 cities as small multiples, and unlike previous range charts, consecutive ranges in the Weather Radials chart are laid out radially to emphasize seasonal variation and the cyclicality of time. More recent radial range charts have presented variations on the Weather Radials design [16, 41, 45], and several radial range charts can be found throughout Lima's recently curated collection of circular charts [44]. Although radial layouts have a long history in the visualization community [24, 26], radial layouts of ranges (such as in the bottom row of Figure 1) have yet to be examined by the research community [29], especially in the context of mobile devices. We thus compare linear and radial range charts for the first time in our experiment.

2.3 Ranges on Mobile Phones

Several existing mobile apps incorporate linear range charts, including weather apps like Dark Sky [23], Weathertron [38], and Weather Line [48]. Linear range charts can also be seen in activity tracking apps such as Azumio's Sleep Time [6] and Garmin's Connect [30], as well as in the sleep tracking features of Apple's Health app [5], in which the quantitative axis corresponds to hours of the day and the range marks indicate the hours slept. Motivated by these apps, we incorporate both temperature and sleep range data in our study.

In the research literature, several applications incorporate range charts. For instance, SleepTight [20] allows a person to identify interesting patterns and inconsistencies in their sleep data, as well as how contextual factors impact their sleep. Recent work by Kay et al. [36] evaluated alternative encodings of bus arrival time distributions on mobile phones that emphasize uncertainty in the span and shape of the distribution, examining how these designs support the comparison of ranges and distributions over time. While we do not consider uncertainty in ranges in our experiment, how uncertainty interacts with the factors considered in our experiment is a promising future direction.

Across the commercial applications mentioned earlier, we observed that the number of ranges displayed at one time can vary; while a week or 10 days of ranges were most common, it is possible to display an entire month of daily sleep duration ranges in the case of Apple's Health app, or 12 temperature ranges corresponding to each month in Weather Line. Among the research applications, SleepTight [20] provided an option to display 7 days, 2 weeks, and 4 weeks of ranges. In a blood pressure tracking application by Chittaro [19], one could pan and zoom in time, resulting in a varying number of ranges displayed on the screen. This variation prompted us to consider the cardinality of ranges as a function of the granularity of time in our experiment.

2.4 Visualization Evaluation Studies

Our work continues the tradition established decades ago [21] of studies that empirically compare visual encoding design choices with human subjects. In recent years, we have seen several experiments compare

alternative visual encodings for time-oriented data, although these have tended to focus on continuous time-series data where graphical marks share a common baseline [2, 28, 33]. An exception is Albers et al.'s study of alternative encodings for aggregated time-series data [4], where they compared a simplified boxplot against seven alternative encodings in the context of six tasks, which included identifying the maxima, minima, and extent of the encoded ranges. In our work, we compare a similar linear range encoding against its radial counterpart with data of varying characteristics and at three granularities of time.

Following the crowdsourced graphical perception work of Heer and Bostock [32], our experiment involved the use of a crowdsourcing platform, which helps to overcome the limitations of controlled lab studies as it provides a diverse and large participant pool [12]. To our knowledge, our work is the first to conduct a visualization evaluation study on participants' mobile phones leveraging a crowd platform.

3 Experiment

To investigate the visualization of ranges over time on mobile phones, we designed an experiment involving four factors: data source, temporal granularity, layout, and task.

3.1 Two Sources of Range Data

Motivated by existing mobile apps profiled earlier and by a desire to maximize external validity, we used real temperature and sleep duration data as sources of ranges over time.

Temperature range data: We used Seattle's daily high and low temperatures for every day of 2015 (from `https://www.wunderground.com`), which we will refer to as *observed ranges*, combined with average high and low temperatures for those same days (from `http://www.intellicast.com`), which we will refer to as *average ranges*.

Sleep duration range data: We used a near-complete year of recorded sleep ranges posted on the /r/datasets subreddit (`https://redd.it/1c1sah`). We used this dataset to generate a semi-synthetic dataset of observed ranges informed by work discussing circadian rhythms and weekday / weekend differences [1]. Our data thus had no missing days or multiple ranges per day resulting from napping or interrupted nighttime sleep. To generate the analog of an average temperature range for sleep data, we considered the medical advice of maintaining a consistent sleep schedule [7] and sleeping for approximately 8 hours each night [22]. We therefore calculated the average bedtime across the dataset, rounded to the nearest half-hour, and designated a recommended sleep time of 8 hours starting from this time.

These two datasets differ in three respects. First, the Seattle temperature ranges exhibit seasonal variation over the course of a year, while the sleep duration ranges exhibit a more periodic pattern, with variation occurring between weekdays and weekends. Second, the average temperature ranges fluctuate seasonally, while the average sleep range is consistent throughout the year. Finally, the quantitative domain of sleep duration ranges is also consistent, with an upper bound of 24 hours (the maximum for a single day), while the quantitative domain of temperature ranges varies considerably across weeks, months, and years. If we consider our set of observed temperature ranges, the domain of temperature ranges for a single week in January is much smaller than the domain of temperature ranges across an entire year (cf. Figure 1), as it is absent of extreme short-term temperature fluctuations.

Both datasets were modified to reduce the impact of skewed distributions and to ensure unique correct responses to the tasks defined in Section 3.5. We removed outlier values and ensured that each week, month, and year contained a single unique maximum and minimum observed range value, replacing duplicate values with other observed values sampled from the same week.

3.2 Three Levels of Temporal Granularity

Existing weather and sleep-tracking mobile apps vary in terms of the number of ranges shown in a single chart, from 7 days of ranges to a range for every day in a month. We were curious as to what the upper bound might be in terms of the cardinality of ranges, or whether it would be possible to display several months of daily sleep duration ranges, as in Boam's "7 Months of Sleep" [11], or a year of daily temperature ranges (e.g., [37, 49]), especially since the pixel resolution of contemporary mobile phones is now larger than the number of days in a year. We therefore decided to consider three granularities of time corresponding to 3 cardinalities of ranges: a **Week** (7 ranges), a **Month** (28–31 ranges), and a **Year** (365 or 366 ranges).

3.3 Two Layouts: Linear and Radial

Inspired by the charts profiled in Section 2.2, we considered both Linear and Radial layouts of time, as shown in Figure 1 and in Table 1.

In a **Linear** range chart, the quantitative scale is orthogonal to the chronological scale; in our case, the chronological scale is horizontal with time proceeding from left to right. For Temperature data (see the top row of Figure 1), cooler temperatures are lower on the screen, as per the convention used in weather reporting. In contrast, the quantitative scale is inverted for Sleep data (cf. the top row of Table 1), with bedtimes for the labeled day appearing toward the top of the screen and waking times for the subsequent morning toward the bottom. While a night's sleep typically spans two calendar days, we adopt a convention used in sleep tracking to attribute sleep beginning before 5am to the previous waking day, rather than to the day in which the waking occurs, so our date labels refer to the day of the bedtime.

In a **Radial** range chart, the quantitative scale emanates from lower values (in our case, cooler temperatures or bedtimes) in the center toward higher values (warmer temperatures or waking times) at the periphery, while the chronological scale is circumferential, proceeding clockwise beginning at 12 o'clock. One potential advantage of a Radial layout is that it emphasizes visual continuity, reinforcing cycles such as a year's seasonal temperature cycles or weekday / weekend sleep routines. On the other hand, one potential disadvantage of a Radial layout is that it may suggest continuity where none exists, such as between the beginning and end of a single month.

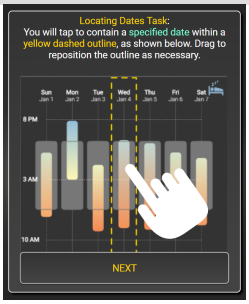
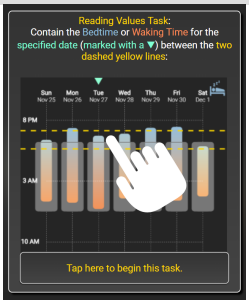
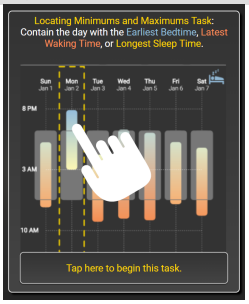
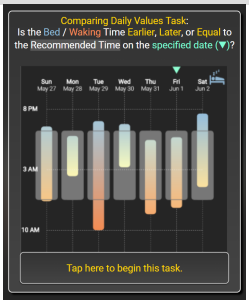
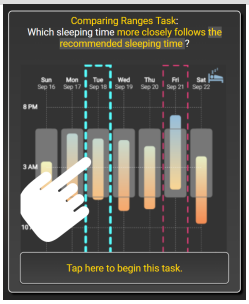
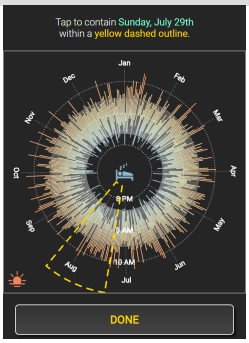
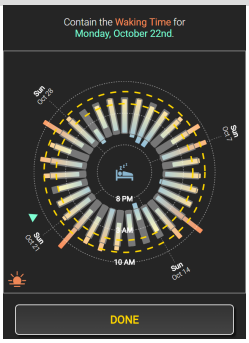
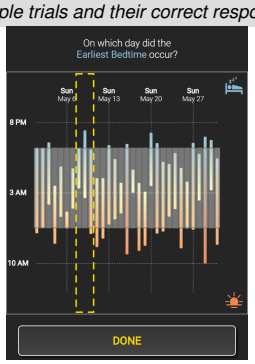
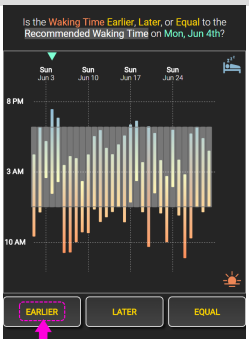
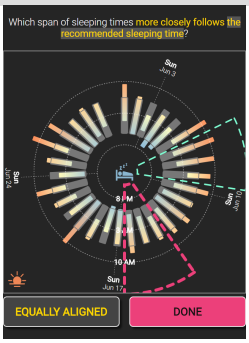
We selected these two layouts to investigate a trade-off between quantitative and chronological scale resolution. As an example, consider an iPhone 6 held in portrait mode, which has a display width of 375 pt and a device-pixel-ratio of 2x, resulting in 750 px. For both of our layouts, we maintained a consistent chart size with a square aspect ratio spanning the entire width of the display, with a margin of 12.5% on all sides of the chart to accommodate axis ticks, leaving 562 px with which to scale the quantitative and chronological domains. With the Linear layout, the quantitative scale extends 562 px vertically while the chronological scale extends 562 px horizontally. In contrast, the Radial layout compresses the quantitative scale to half of the chart width (281 px), since the scale emanates from the center toward the periphery. Accordingly, all of the range marks in a Radial layout are half the length of their Linear counterparts. Meanwhile, the chronological scale extends circumferentially in a Radial layout, spanning 1766 px at the periphery. Thus, we expected that performance may suffer with the Radial layout when discriminating between quantitative values and lower values in particular. Conversely, we expected that one might benefit from the increased chronological resolution of a Radial layout when discriminating between chronological values. Furthermore, the gap between adjacent range marks is greater with a Radial layout (compare the top and bottom row of Figure 1), and we were curious about how this affects value discrimination at the different granularities of time.

3.4 Range Encoding

Our Linear and Radial range charts both feature the same visual encoding choices for observed and average range marks.

The observed range marks incorporate a redundant color encoding of the ranges inspired by the "Weather Radials" chart [37], which assigned a single color value to each range based on the day's average temperature value. While we expected that a redundant color encoding might be beneficial to viewers, we noted that there is no analogously meaningful "average sleep value" for a sleep duration range. Therefore, instead of a single color value for each range, our implementation applies a unique color gradient to each range via a LAB interpolation through a 3-class diverging red-yellow-blue scale [17] mapped to the quantitative domain, following Bremer's method [15]. As a result,

Table 1. Our experiment included five experimental tasks, presented to participants in this order, from relatively simple to complex. This table illustrates these tasks using Sleep range data. Each task was preceded by an introductory instruction (top row). The second row shows an example trial for each task. The table also indicates the possible response values and the number of trials for each task. To complement this table, we provide a video of representative experimental trials as supplemental material at `https://github.com/Microsoft/rangesonmobile`.

Task 1: **Locate Date**	Task 2: **Read Value**	Task 3: **Locate Min / Max**	Task 4: **Compare Values**	Task 5: **Compare Ranges**
Introductory instruction shown to participants:				
				
Example trials and their correct responses:				
				
Possible responses:				
Given that the span of the *response indicator* (centered around the touch point) is 1, 3, and 31 days at the Week, Month, and Year levels of granularity, respectively, does the *response indicator* contain the specified date?	Given that the span of the *response indicator* is 10% of the quantitative domain, does the *response indicator* contain the specified target value for the indicated date?	Given that the span of the *response indicator* is 1, 3, and 31 days at the Week, Month, and Year levels of granularity, respectively, does the *response indicator* contain the date having the specified min or max value?	3 possible responses: the specified target value on the indicated day is [earlier (Sleep) / cooler (Temperature)], [later / warmer], or equal to the average value. (Pink arrow added above to show correct response.)	3 possible responses: either the range(s) contained by the pink *response indicator* or those contained by the cyan *response indicator* are more closely aligned to their corresponding average ranges, or they are equally aligned.
Number of test trials across levels of granularity:				
Week: 2	Week: 2 x (Start, End)	Week: 2 x (Start, End, Span)	Week: 2 x (Start, End)	Week: 2
Month: 2	Month: 2 x (Start, End)	Month: 2 x (Start, End, Span)	Month: 2 x (Start, End)	Month: 2
Year: 2	Year: – *(omitted)*	Year: 1 x (Start, End)	Year: – *(omitted)*	Year: 2
Total: 6 Linear + 6 Radial	Total: 8 Linear + 8 Radial	Total: 14 Linear + 14 Radial	Total: 8 Linear + 8 Radial	Total: 6 Linear + 6 Radial

the start of ranges appear bluer to reinforce the semantics of cooler temperatures or nighttime, while the end of ranges appear redder to reinforce warmer temperatures or daytime.

Each observed range mark is superimposed over a wider grey rectangular mark encoding the average range for that day. We also overlay grey line marks corresponding to low and high average values above the observed range, allowing the silhouette of average ranges to remain visible at higher granularities of time (e.g., Figure 1-right).

3.5 Tasks

We selected five tasks (indicated in Table 1) that span varying levels of difficulty: *Locate Date*, *Read Value* (for an indicated date), *Locate Min / Max* (range value), *Compare Values* (for an indicated date), and *Compare Ranges* (in two indicated regions). Each of these tasks involves position judgments along either a common chronological or quantitative scale, where the marks do not share a common baseline [29]. Despite our consideration of Radial layout, none of our selected tasks involve angular judgments, such as estimating the time elapsed between two ranges. We derived these tasks from Brehmer and Munzner's task typology [14], by considering the various combinations of two *Search* actions (*Lookup* and *Locate*) with two *Query* actions (*Identify* and *Compare*) in the context of ranges over time.

3.6 Research Questions

Prior to data collection we framed three research questions that we intended to ask in the context of both sources of range data:

Q1 / Layout: How does layout (Linear or Radial) affect performance across the five tasks?

Q2 / Granularity: How does the temporal granularity (and thus the cardinality of ranges) affect performance across the five tasks?

Q3 / Target range value: For tasks that ask participants to attend to either the Start or End values of ranges (*Read Value*, *Locate Min / Max*, *Compare Values*), do participants benefit from increased scale resolution at the periphery of a Radial layout, and conversely does performance degrade from smaller marks and a reduced chronological resolution at the center?

Note that we do not include a research question regarding the two data sources, as an analysis of the differences between these two groups would not be appropriate. This is due to confounding factors that include differences in range value distributions (as described in Section 3.1), subtle differences in task instruction wording reflecting different data semantics, and an inversion of the quantitative scale for the Linear representation (as described in Section 3.3).

3.7 Experiment Design

We conducted a mixed-design study with repeated measures as an online experiment. Data source was a between-subjects factor while task, granularity, and layout were within-subjects factors. The order of tasks was fixed as numbered in Table 1, while the order of layouts and granularities was randomized for each participant and for each task.

Altogether, the experiment required approximately 30 minutes to complete, including an introductory tutorial and a subjective response survey at its conclusion. The introductory tutorial asked participants to hold their phone in portrait mode with their non-dominant hand or to lay their phone on a flat surface, to respond to tasks by tapping with the index finger of their dominant hand, and to take breaks between tasks as necessary. These instructions were included to promote consistency in participant behavior and to reduce the impact of fatigue incurred by holding a phone for half an hour. We also wanted to discourage participants from responding with the thumb of the dominant hand while holding their phone, as larger phones can impede movement of the thumb, especially for people with smaller hands. The introduction proceeded to incrementally introduce annotated versions of the visual encodings used in the experiment. Each of the five tasks was also preceded by a brief task-specific introductory instruction indicating how to complete each task (top row of Table 1).

Table 1 also indicates the number of trials per task. For each task / granularity / layout combination, participants first completed a practice trial and received feedback regarding the correctness of their response; if incorrect, they could try again until they responded correctly, and we highlighted the correct response after two failed attempts. Participants then completed between 2 and 6 test trials for each granularity and layout, depending on the task: the *Read Value*, *Locate Min / Max*, and *Compare Values* tasks each included 2 trials that asked about the start values of ranges, as well as 2 trials that asked about the end values of ranges; the *Locate Min / Max* task additionally included 2 trials at the Week and Month levels of granularity which asked participants to locate the range with the largest span. We omitted the Year level of granularity for the *Read Value* and *Compare Values* tasks, as pilot participants reported these trials to be exceedingly difficult; we similarly reduced the number of *Locate Min / Max* trials at the Year level. For tasks other than *Locate Min / Max*, we selected target days at random, and we ensured that any single Week or Month of the range dataset was randomly selected in at most one Week or Month trial per task.

We also interspersed 6 additional trials of the relatively easy *Locate Date* task at the Week level of granularity at various points throughout the latter four tasks, which we used to test the participants' attention and their ability to respond to changing instructions. When a participant responded incorrectly to any of these "quality control" trials, we used this as an indicator of inattention to the task and excluded the participant's data from any subsequent analysis. Taking into account all task / granularity / layout combinations and excluding practice and quality control trials, participants completed a total of 84 trials.

3.8 Metrics

Completion Time: In all tasks, a dialog indicating the trial instruction preceded each trial; tapping on it revealed the corresponding range chart and triggered the start of a timer.

In the *Locate Date* and *Locate Min / Max* tasks, we asked participants to tap on the chart to contain a specified target value inside of a rectangle or wedge marked with a yellow dashed outline centered around their touch point (see Table 1), which we will refer to as a *response indicator*. We devised the response indicator to counteract the *fat finger problem*, as our focus in this experiment was not on fine-grained touch-based selection. While the geometry and size of the response indicator differed between Radial and Linear layouts relative to the physical chart dimensions, we ensured that the response indicator was sufficiently large, spanning the equivalent domain of possible response values along the chronological or quantitative scale. In the *Locate Date*, *Locate Min / Max*, and *Compare Ranges* tasks, the size of the response indicator varied according to the temporal granularity of the data: at the Week level of granularity, the response indicator spanned a single day, while at the Month and Year levels of granularity, the response indicator spanned 3 days and 31 days, respectively. Since the response indicator was wider than a single day at the Month and Year levels of granularity, participants were told that the target does "not have to be exactly in the middle." In the *Read Value* task, the response indicator was a rectangle or pair of concentric rings spanning 10% of the quantitative domain (see the second column of Table 1).

Once the response indicator appeared, the "Done" button at the bottom of the screen became enabled. The trial Completion Time is thus the time until the participant tapped on this button. The *Compare Values* task provided a fixed set of three responses, so the trial Completion Time was the time until the participant pressed any one of these. Finally, the Completion Time in the *Compare Ranges* task was either the time until the participant tapped on the "Done" button, which became activated after tapping on either the pink or the cyan response indicator in the chart, or until the participant tapped on the "Equally Aligned" button.

Error Rate: In the first three tasks, we classified a participant's response as an error if the specified target value fell outside of the response indicator when they tapped on "Done." In the *Compare Values* and *Compare Ranges* tasks, only one of the three possible responses was correct in each trial.

Subjective Responses: The concluding survey asked participants to select their preferred layout and to rate their confidence with both layouts from 1 (low) to 5 (high) at each level of granularity.

3.9 Participants

We initially recruited 13 current and former colleagues for a pilot experiment; an experimenter personally observed four pilot experiment sessions and solicited feedback regarding the usability of the experimental application and the difficulty of the tasks. The remaining pilot participants were remote and provided feedback via email.

Since those in our first pilot experiment were generally familiar with HCI and / or visualization, we conducted a second pilot experiment via Amazon Mechanical Turk with 12 participants. This second pilot gave us an impression of how long the experiment would take with crowd workers and allowed us to calibrate the quality control trials.

Satisfied with the results of the second pilot, we recruited MTurk workers with a planned sample size of $N = 100$, randomly assigning 50 to a Temperature group and 50 to a Sleep group. We limited our recruitment to "masters" level workers with HIT approval ratings of 99% or higher and to those from the United States. To attain additional consistency across participants, we specified that workers use a smartphone running iOS 9 or greater or Android 5 or greater, that they use either the Chrome or Safari mobile browser, and that they have adequate battery power and a stable WiFi connection prior to beginning the study. As our experiment took approximately 30 minutes to complete, we paid each crowd worker $4 USD, being slightly higher than the current federal minimum wage in the USA. Upon completion of the experiment, we asked participants to copy a completion code provided by our application into the MTurk interface; if no code was provided after 60 minutes of starting the experiment, we considered the session to be abandoned and it was re-assigned to another crowd worker. Finally, we disallowed workers from participating in the experiment multiple times, and to this end our application used a browser cookie to prevent the application from loading if it detected a repeat visitor.

3.10 Implementation and Deployment

Our experimental software is a Node.js application [35], which we deployed as an Azure web app [46] at `https://aka.ms/ranges`, which allowed us to log responses to our experimental tasks as custom events with Azure's Application Insights tool. We used D3.js [13] (v4) to visualize the range data. The web app can only be viewed from a mobile phone held in portrait mode, and it is compatible with recent versions of mobile web browsers such as Chrome and Safari. The source code of our application is available under a MIT open source license at `https://github.com/Microsoft/rangesonmobile`.

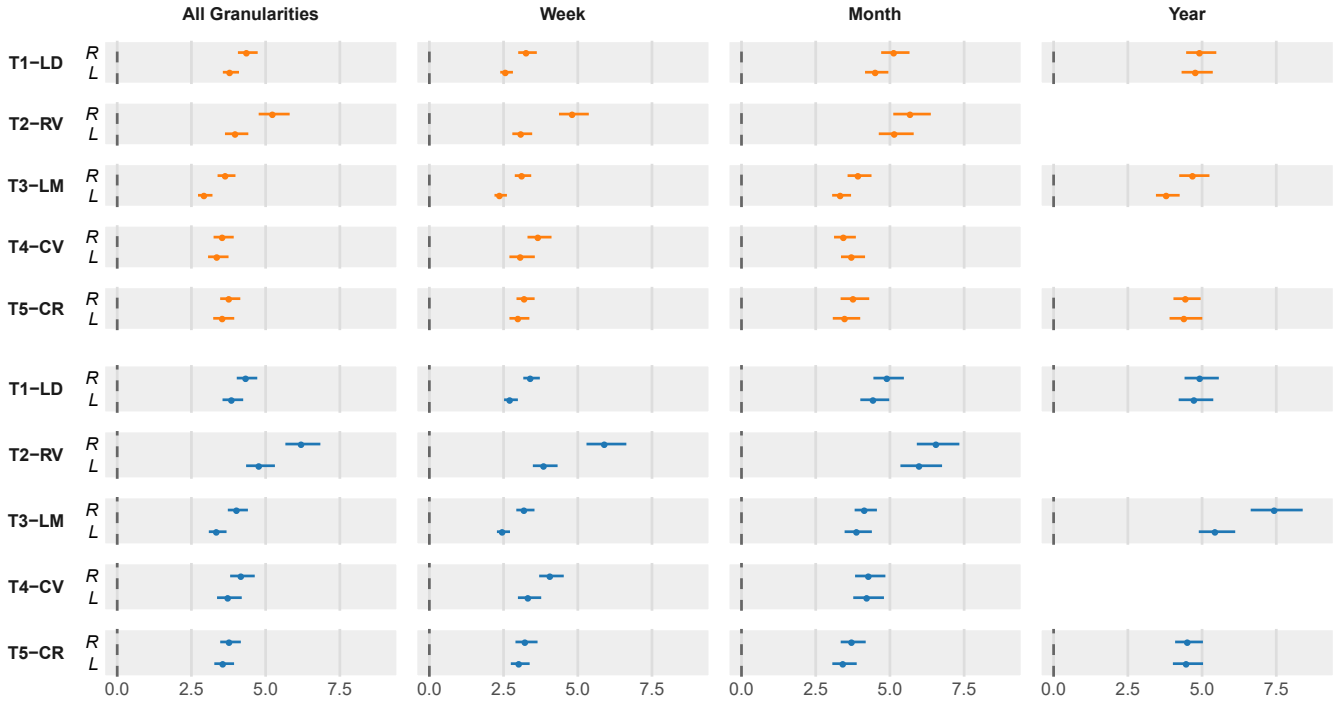


Fig. 2. Mean Completion Times in seconds for tasks 1-5 with *Radial* and *Linear* layouts. Task abbreviations: T1-LD = *Locate Date*, T2-RV = *Read Value*, T3-LM = *Locate Min / Max*, T4-CV = *Compare Values*, T5-CR = *Compare Ranges*. ● = Temperature; ● = Sleep. As indicated in Table 1, T2 and T4 did not include trials at the Year level. Error bars are 95% CIs; for guidance with respect to interpreting overlaps in CIs, refer to [40].

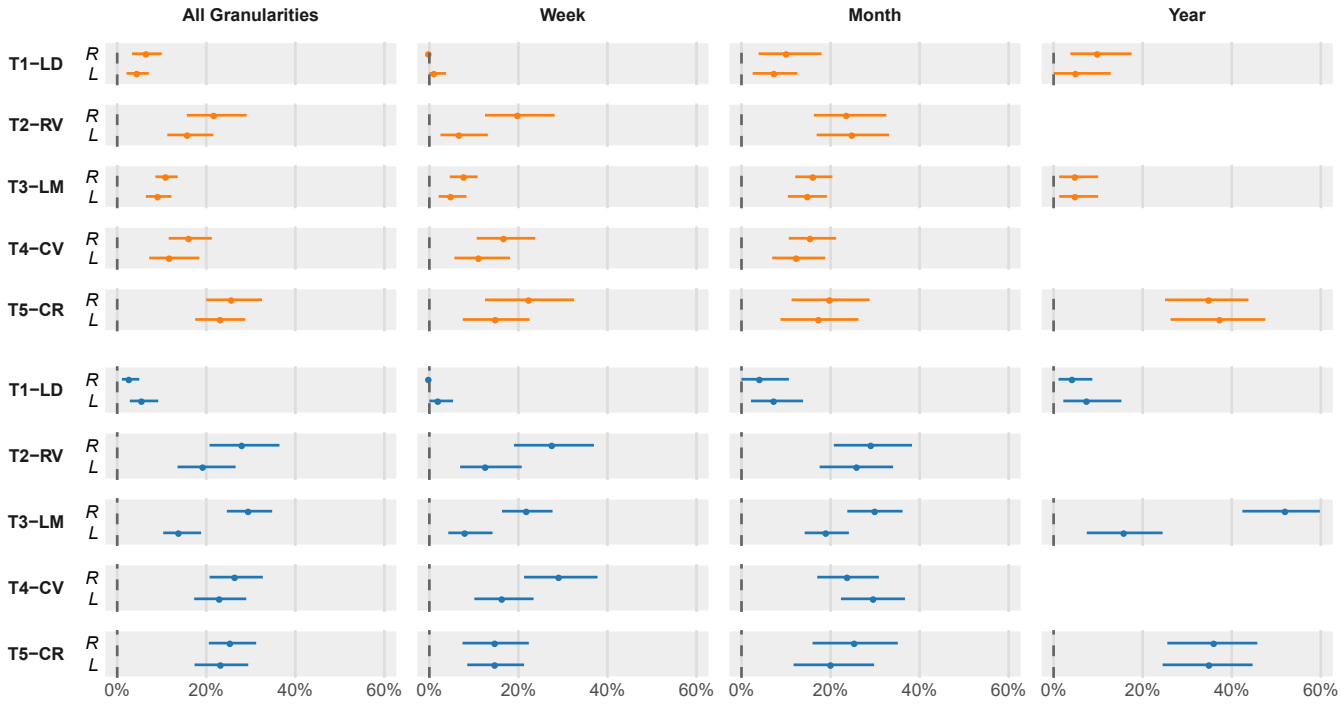


Fig. 3. Error Rates for tasks 1-5 with *Radial* and *Linear* layouts. ● = Temperature; ● = Sleep.

4 Results

We analyze, report, and interpret all of our inferential statistics using interval estimation [25]. Our experiment data and our analyses are available alongside the application source code in the same repository.

We report the sample mean for task / granularity / layout combination according to the metrics defined in Section 3.8. We also report 95% confidence intervals (CIs) indicating the range of plausible values for the population mean. For Completion Time, we use t-statistic confidence intervals, while we use BCa bootstrap confidence intervals for Error Rate and participants' self-reported Confidence [9]. We log transformed participants' Completion Times to correct for positive skewness and we present anti-logged geometric mean Completion Times [25, 52].

We excluded data from six participants from the Temperature group as well as from two participants from the Sleep group for failing to respond correctly to our quality control trials. We excluded data from another four participants in the Temperature group and one from the

Sleep group for reasons of noncompliance: for providing an incorrect completion code to circumvent the 60-minute timeout or for participating from a non-English-speaking country. As a result, our final participant counts were $N(Temperature) = 40$ and $N(Sleep) = 47$. These participants used Chrome or Safari on mobile phones running Android (5.0–8.1) or iOS (10.0–11.2) with resolutions (pt) ranging from 320w x 445h to 424w x 674h. We also excluded practice and quality control trials, as well as outlier trials with Completion Times longer than 3 standard deviations from the mean (after log transformation), leaving 3,926 Sleep group trials and 3,337 Temperature group trials for our confidence interval calculations.

4.1 Overview of Results

Completion Time: Figure 2 shows mean Completion Times, with tasks as rows, granularities as columns, and the two data groups distinguished using color (● = Temperature; ● = Sleep). Averaged across all granularities, participants completed trials between 2.8 and 5.4 seconds with Linear layouts and between 3.3 and 6.9 seconds with Radial ones (see the first column of Figure 2). Participants' Completion Times grew as the granularity increased. While participants completed trials of most tasks in about 4 seconds, Completion Times tended to be higher in the *Read Value* and *Locate Min / Max* tasks, particularly at the Month level of granularity for the former and at the Year level for the latter.

Error Rate: Figure 3 shows that averaged across granularities, the Error Rate varied considerably by task and was between 2% and 30% with Linear layouts and between 1% and 37% with Radial ones. Within individual tasks, we saw more errors in the *Compare Ranges* task at the Year level, and while the two groups generally incurred a similar number of errors, those in the Sleep group committed more errors in the *Locate Min / Max* and *Compare Values* tasks.

Subjective responses: As shown in Figure 4 and Figure 5, more participants preferred a Linear layout to a Radial one, and they were more confident using a Linear layout, though the proportion of participants who preferred the Radial layout grew as granularity increased, particularly among those in the Sleep group; at the granularity of a Year, these participants were about evenly split between the two layouts. Unsurprisingly, participants were less confident as the granularity increased.

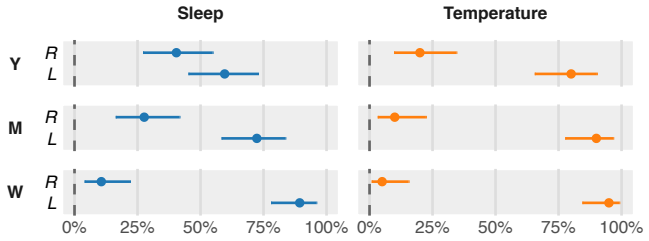


Fig. 4. Proportion of participants who prefer either a *Radial* or *Linear* layout at each granularity (*Week* / *Month* / *Year*).

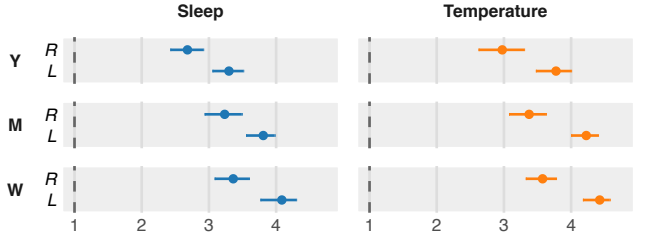


Fig. 5. Participants' reported confidence from 1 (low) to 5 (high) with *Radial* and *Linear* layouts at each granularity (*Week* / *Month* / *Year*).

4.2 Result Analyses

We now report effect sizes relating to our research questions from Section 3.6. We planned all analyses in this section before collecting data. For Completion Time, we report pair-wise comparisons of means; since these comparisons are differences in log-transformed values, we present them anti-logged as ratios between geometric means [9]. For Error Rate, we report pair-wise comparisons as differences in means. As before, we also report 95% CIs.

Q1 / Layout: Figure 6 shows ratios between mean Completion Times, indicating that participants in both groups were up to slower with Radial layouts in the first three tasks. This difference was more pronounced with the *Read Value* task (20%–42% slower) and the *Locate Min / Max* task (13%–30% slower). On the other hand, participants were not slower with Radial layouts when performing the two comparison tasks, with the exception of those in the Sleep group performing the *Compare Values* task, where they were 2%–21% slower with a Radial layout. Unexpectedly, ratios in Completion Time between the two layouts shrank as the granularity increased (Figure 7).

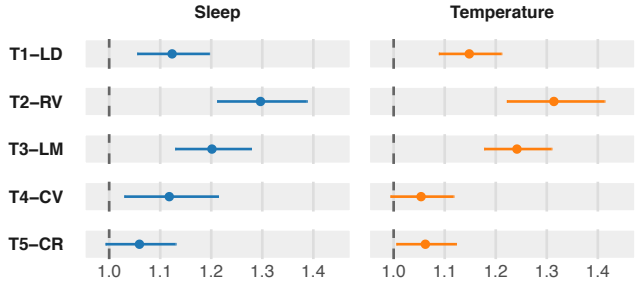


Fig. 6. *Radial* / *Linear* Completion Time ratios by task. In cases where a CI intersects the dashed line (a ratio of 1), we interpret this as inconclusive evidence for an effect [9].

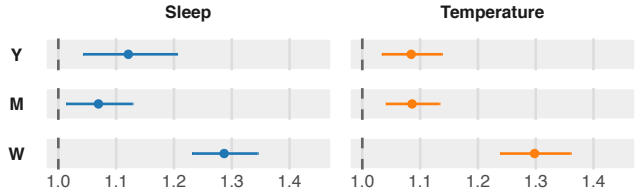


Fig. 7. *Radial* / *Linear* Completion Time ratios by granularity.

Figure 8 shows that we have inconclusive evidence to suggest that either layout incurred more errors for Temperature group participants in any single task. In contrast, with a Radial layout, Sleep participants committed 1%–8% fewer errors in the *Locate Date* task, 2%–16% more errors in the *Read Value* task, and 11%–20% more errors in the *Locate Min / Max* task. We also have little evidence for a difference in Error Rate among Sleep group participants for the comparison tasks.

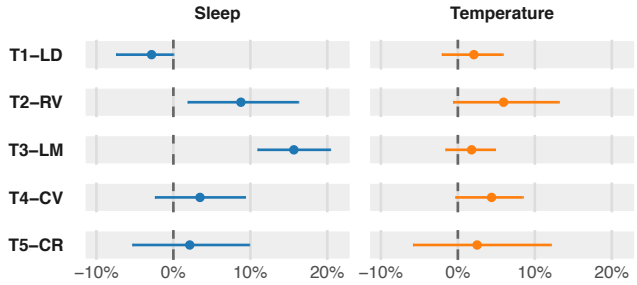


Fig. 8. *Radial* − *Linear* Error Rate differences by task. In cases where a CI intersects the dashed line (a difference of 0%), we interpret this as inconclusive evidence for an effect.

Q2 / Granularity: Figure 9 shows that participants were typically slower to complete tasks with a Month or Year of ranges than with Week of ranges, and in all but the *Locate Date* task, participants were slower with a Year of ranges than with Month of ranges. In *Locate Min / Max* task, it is worth noting that those in the Sleep group took more than twice as long (107%–149%) in Year trials than in Week trials, which was a much larger difference than what we saw from the Temperature group (43%–71% slower).

Figure 10 shows that Error Rates were higher from Month to Week in the first three tasks and that increases in granularity beyond that did not necessarily result in a higher Error Rate. For instance, in the *Locate Min / Max* task, the Error Rate at the Year level was 5%–15% lower than at the Month level among those in the Temperature group. In the Sleep group, differences in Error Rate were similar to those of the Temperature group except in the *Locate Min / Max* task, where the Error Rate was 3%–15% higher from Year to Month.

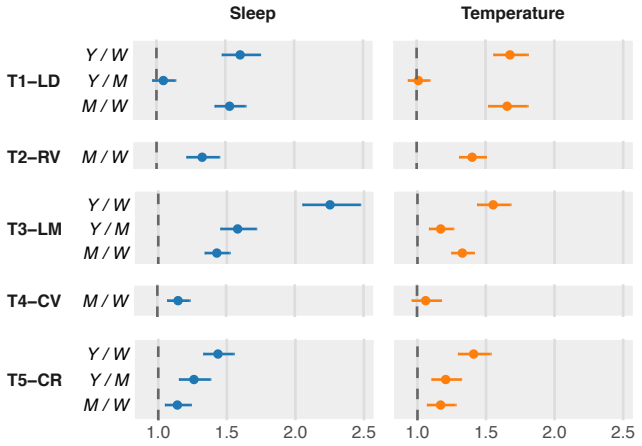


Fig. 9. *Pairwise Year, Month, and Week* Completion Time ratios.

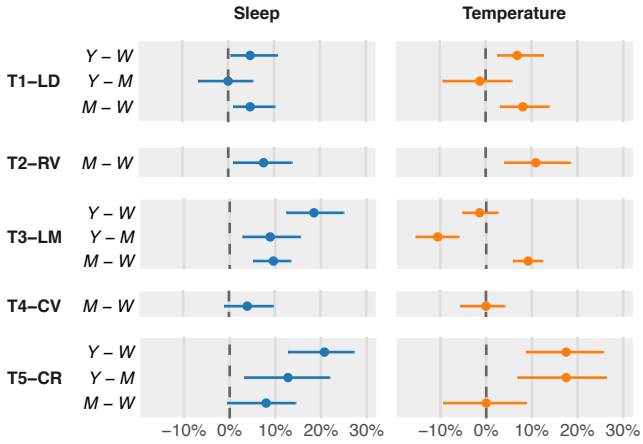


Fig. 10. *Pairwise Year, Month, and Week* Error Rate differences.

Q3 / Target range value: Figure 11 shows that participants were slower with Radial layouts regardless of whether the task asked them to attend to the Start or End value of a range, except in the *Compare Values* task. Additionally, we found no evidence of a difference in Error Rate between the layouts when we faceted the results by target value (Figure 12), at least for the Temperature group.

However, we did find Error Rate differences in the Sleep group, where Radial layouts incurred more errors in the *Locate Min / Max* task regardless of target value, as well as 1%–10% more errors when reading Start values in the *Read Value* task.

5 Discussion

Considering the analyses presented in the preceding section, we now discuss design implications for visualizing ranges on mobile phones and opportunities for future research.

5.1 Implications for Design

Our chart stimuli and response interactions were intended for a mobile phone held in portrait mode and are not well-suited for a large display, such as a desktop. Therefore, our findings and the discussion in this section should not be used to inform the design of charts depicting ranges over time on devices other than mobile phones.

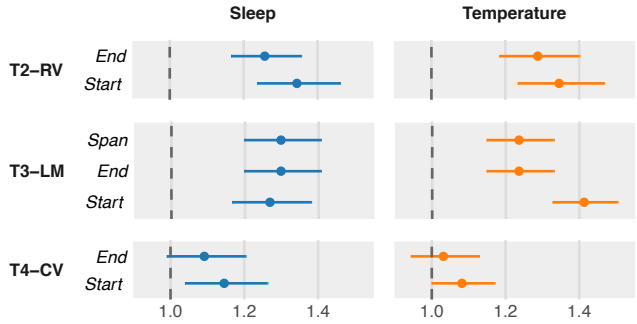


Fig. 11. *Radial* / *Linear* Completion Time ratios by target range value.

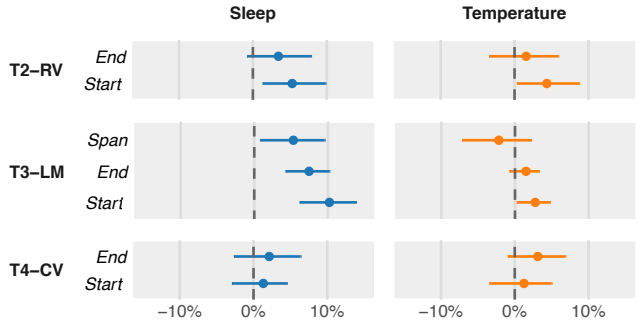


Fig. 12. *Radial* − *Linear* Error Rate differences by target range value.

How to decide between a Radial or Linear layout. There has been a resurgent interest in charts with circular or Radial layouts, both from the research community (e.g., [2, 28, 29]) and from the visualization practitioner community, as such charts are often praised for their aesthetic qualities [44]. For instance, Kekeritz's "Weather Radials" [37] received notable press coverage and an Information is Beautiful award in 2014. Although we have seen many circular or Radial layouts in charts designed for print and larger displays, we were curious about the prospect of a Radial layout for ranges on mobile phone screens.

Despite their potential aesthetic appeal, our participants did not prefer range charts with a Radial layout, opting instead for a Linear layout, one that they felt more confident in using. Our participants' subjective responses are reinforced by the finding that they were generally slower with a Radial layout, but this difference in speed is less pronounced at higher granularities of time and was not apparent when performing comparison tasks. Contrary to our expectations, we were most surprised to learn that despite a greater visual emphasis on higher values at the periphery of a Radial layout, participants still took longer to *Read Values* or *Locate Min / Max* values with this layout, and those in the Sleep group committed more errors with this layout when they were asked to locate the day with the latest waking time.

In terms of design guidance, for use cases that involve comparing observed range values against average range values, we would expect similar performance from either a Linear or Radial layout. Example use cases would include determining if a month's observed temperature ranges deviated from average temperatures, or determining if one's sleep schedule has drifted from an idealized schedule. If the use case involves reliably *Locating Min / Max* values, we recommend against the use of a Radial layout. However, it is still possible that a Radial layout provides an advantage in tasks other than those we tested, such as identifying trends or deviations from seasonal patterns over the course of several years (e.g., Hawkins' superimposed radial line graphs in *Climate Spirals* [31]), which we leave as a question for future work.

How many ranges can you show on a mobile phone display? The weather and sleep-tracking apps that we profiled in Section 2.3 varied in terms of how many ranges they displayed on a single screen, from 7 to 31. Despite a small display size, we were curious as to whether an entire Year of daily ranges could be shown on a mobile display and still be put to practical use. For instance, a person might want to reflect

on their sleep habits over the course of a year, or examine the annual temperature variation of a city to determine the best time to visit.

Considering our results, the answer to this question depends in part on the task and in part on the source of range data. For participants who saw Temperature ranges, a Year of ranges did not pose great difficulty, and when they were asked to *Locate Min / Max* values, their worst performance was not with a Year of ranges but with a Month of ranges, which may simply be an unconventional time window for consuming weather information. These participants also remained fairly confident with a Year of ranges, at least relative to those who saw Sleep data, whose performance generally worsened as the granularity increased.

Also of note are the results of the *Compare Ranges* task, in which a Year of ranges incurred a high number of errors, regardless of the source of the data or the layout of the chart.

An alternative to showing a full Year of daily ranges would be to aggregate them into weekly or monthly ranges, like how the WeatherLine app shows an average temperature range for each month [48]. The disadvantage of aggregation is that the presence and effect of extreme values is not immediately apparent. Thus we would need to consider the encoding of outliers, adopt a boxplot-like design, and / or permit a drill down interaction from a Year to a Month or Week. A second alternative would be to show a year of daily ranges only when the phone is held in landscape mode, although this would preclude the use of a Radial layout, as it requires a square aspect ratio.

Congruence with data and task. Another way to frame our research questions is to ask which combination of layout and granularity is congruent with the combination of data and task. In particular, we noted throughout this paper that a Radial layout reinforces the cyclicality of time. However, not all cycles are of interest for all sources of range data. For instance, a week is a cycle defined by cultural convention, one in which many people adopt a "weekly routine" with a desired sleep schedule, so comparing sleep durations to this schedule is congruent with a Radial layout at this granularity. In contrast, the weather does not follow a weekly cycle but an annual one, and thus a Radial layout may only be justifiable for comparison tasks at this larger granularity. Similarly, a month is unlikely to be a meaningful cycle length for either sleep routines or temperatures, whereas it might be meaningful in the context of lunar or tidal cycles.

5.2 Limitations & Future Work

Any comparative experiment such as ours involves tradeoffs between external validity and control over possible confounds. Our crowdsourcing approach allowed us to recruit many participants who could perform the experiment using their own mobile phones without an experimenter present to observe (and impact) their performance. Yet, we had no way to control the context in which participants performed our experiment. Similarly, our choice of datasets may not be representative of all possible distributions of temperature and sleep range values or of other sources of range data.

Despite these limitations, considering our manipulation of the datasets described in Section 3.1, two pilot experiments, and our calibration of practice and quality control trials, we believe that our results are representative of typical performance. Going forward, we want to account for other factors that we could not accommodate in this study, specifically focusing on the following avenues for future work.

The role of range semantics and redundant encoding. Our experimental stimuli and task instructions repeatedly conveyed the semantics of the ranges, featuring temperature- and sleep-related iconography and word choices. Similarly, the redundant color gradient encoding for observed ranges reinforced the warm / cool continuum with Temperature data and the nighttime / sunrise dichotomy with Sleep data. An interesting direction for future work could involve removing these semantic cues in an effort to determine how they affect task performance, or using other sources of range data, such as the daily range of a stock's selling price or the daily range of a building's energy demand.

A personal relationship with ranges. We can assume that most people consume temperature information more often than information pertaining to when they sleep. Moreover, people such as our participants are likely to have expectations with regards to seasonal variation in temperature, whereas our participants had no personal connection to the sleep ranges that we presented to them. It would therefore be worthwhile to repeat our experiment with self-identified "quantified selfers" examining their own data. Many quantified selfers track their sleep habits and thus would have expectations with regards to periodic patterns across ranges as well as explanations for deviations from an average range. Similarly, we could show future study participants weather data from where they live; in our experiment, we used Seattle's daily high and low temperatures, and our application logs revealed that none of the Temperature group participants were located in the Pacific Northwest, so we are unable to determine if a lived experience of these ranges would have affected task performance.

Beyond a single mobile visualization study. While designing and piloting our experimental application, we iteratively refined response interactions for the five experimental tasks so as to be compatible with a mobile phone and with a crowdsourced deployment, simplifying the interactions as much as possible. In the first three tasks, the only interaction provided by the chart interface is used to complete the task: tapping on the chart to position the *response indicator* around a target value, designed to reduce the impact of the fat finger problem. We also asked participants to respond to tasks by tapping with their index finger, since larger phones may disadvantage reaching with one's thumb while holding the phone in the same hand. The two comparison tasks featured a relatively simpler response interaction: each trial asked participants to tap on one of three options. It would be interesting to conduct comparative evaluations of alternative mobile interaction design choices for different combinations of task and datatype leveraging a crowdsourced approach, particularly with response interactions that are less impacted by how the phone is held or by which finger(s) provide touch input. This might help us attain more separable estimates of perceptual difficulty and response difficulty. More generally, it is a promising avenue for future work to develop a methodology or framework to guide researchers with respect to conducting crowdsourced studies of (interactive) data visualization on mobile devices. Finally, experimenters should also investigate ways to increase instruction compliance and response quality by requesting access to the phone's orientation, microphone, WiFi, and ambient light sensors, so that we might discard and replace trials where a detectable change in the environment occurs.

6 Conclusion

We reported results from a crowdsourced visualization experiment, the first to be conducted exclusively on mobile phones. Our experiment focused on ranges over time, a type of data that is often consumed via a mobile phone. Our goal was to identify limitations in terms of how many ranges could feasibly be displayed on a small screen, and to compare participants' task performance when using either a Linear or Radial layout of range marks. In terms of the cardinality of ranges, our analyses revealed limitations that vary by task and data source; although temperature ranges and sleep duration ranges share the same data abstraction of ranges over time, we observed several instances where performance differed between the two sources of range data. With respect to layout, participants generally performed tasks more quickly with a Linear layout, though both layouts generally incurred a similar number of errors. In tasks that involve comparing observed and average range values, we expect people to perform similarly with both layouts. Our results motivate several directions for future work in the context of mobile data visualization, including a consideration of effective ways to aggregate ranges for display on mobile phones, as well as further studies to compare mobile visualization design choices.

Acknowledgments

We thank Pierre Dragicevic for his suggestions regarding result analyses, Ken Hinckley, Catherine Plaisant, and Lonni Besançon for their comments on the paper, as well as our pilot participants for their feedback on the experiment application and procedure.

REFERENCES

[1] S. Abdullah, M. Matthews, E. L. Murnane, G. Gay, and T. Choudhury. Towards circadian computing: "Early to bed and early to rise" makes some of us unhealthy and sleep deprived. In *Proceedings of the ACM Conference on Pervasive and Ubiquitous Computing (UbiComp)*, 2014. http://doi.acm.org/10.1145/2632048.2632100.

[2] M. Adnan, M. Just, and L. Baillie. Investigating time series visualisations to improve the user experience. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 5444–5455, 2016. http://doi.acm.org/10.1145/2858036.2858300.

[3] R. Al-Tarawneh, S. R. Humayoun, and A.-K. Al-Jaafreh. Towards optimizing the sunburst visualization for smart mobile devices. In *Proceedings of the IFIP International Conference on Human-Computer Interaction (INTERACT)*, vol. 22, 2015. https://aka.ms/altarawneh2015.

[4] D. Albers Szafir, M. Correll, and M. Gleicher. Task-driven evaluation of aggregation in time series visualization. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 551–560, 2014. http://doi.acm.org/10.1145/2556288.2557200.

[5] Apple. Health, 2018. https://www.apple.com/ios/health.

[6] Azumio Inc. Sleep Time, 2016. http://www.azumio.com/s/sleeptime/index.html.

[7] L. K. Barber, D. C. Munz, P. G. Bagsby, and E. D. Powell. Sleep consistency and sufficiency: Are both necessary for less psychological strain? *Stress and Health*, 26(3):186–193, 2010. https://doi.org/10.1002/smi.1292.

[8] D. Baur, B. Lee, and S. Carpendale. Touchwave: Kinetic multi-touch manipulation for hierarchical stacked graphs. In *Proceedings of the ACM Conference on Interactive Tabletops and Surfaces (ITS)*, pp. 255–264, 2012. https://doi.org/10.1145/2396636.2396675.

[9] L. Besançon and P. Dragicevic. La différence significative entre valeurs p et intervalles de confiance (The significant difference between p-values and confidence intervals). In *Conférence Francophone sur l'Interaction Homme-Machine*, 2017. https://hal.inria.fr/hal-01562281.

[10] K. Blumenstein, C. Niederer, M. Wagner, G. Schmiedl, A. Rind, and W. Aigner. Evaluating information visualization on mobile devices: Gaps and challenges in the empirical evaluation design space. In *Proceedings of the ACM Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization (BELIV)*, pp. 125–132, 2016. http://doi.acm.org/10.1145/2993901.2993906.

[11] E. Boam. 7 Months of Sleep, 2014. https://ericboam.com/7-Months-of-Sleep.

[12] R. Borgo, B. Lee, B. Bach, S. Fabrikant, R. Jianu, A. Kerren, S. Kobourov, F. Mcgee, L. Micallef, T. Landesberger, K. Ballweg, S. Diehl, P. Simonetto, and M. Zhou. Crowdsourcing for information visualization: Promises and pitfalls. In D. Archambault, H. Purchase, and T. Hossfeld, eds., *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*, pp. 96–138. Springer, 2017. http://www.springer.com/gp/book/9783319664347.

[13] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-driven documents. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 17(12):2301–2309, 2011. http://doi.org/10.1109/TVCG.2011.185.

[14] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 19(12):2376–2385, 2013. https://doi.org/10.1109/TVCG.2013.124.

[15] N. Bremer. Boost D3.js charts with SVG gradients. *Creative Bloq*, 2016. https://aka.ms/bremer-svg-gradients.

[16] N. Bremer. Linear SVG Gradient - Weather Radial, 2016. https://aka.ms/bremer-weather-radial.

[17] C. Brewer, M. Harrower, B. Sheesley, A. Woodruff, and D. Heyman. ColorBrewer 2.0: 3-class RdYlBu scale, 2013. http://colorbrewer2.org.

[18] Y. Chen. Visualizing large time-series data on very small screens. In *Short Paper Proceedings of the Eurographics Conference on Visualization (EuroVis)*, 2017. http://doi.org/10.2312/eurovisshort.20171130.

[19] L. Chittaro. Visualization of patient data at different temporal granularities on mobile devices. In *Proceedings of the ACM Conference on Advanced Visual Interfaces (AVI)*, pp. 484–487, 2006. http://doi.acm.org/10.1145/1133265.1133364.

[20] E. K. Choe, B. Lee, M. Kay, W. Pratt, and J. A. Kientz. SleepTight: Low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the ACM Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pp. 121–132, 2015. http://doi.acm.org/10.1145/2750858.2804266.

[21] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. http://doi.org/10.1080/01621459.1984.10478080.

[22] Consensus Conference Panel, N. F. Watson, M. S. Badr, G. Belenky, D. L. Bliwise, O. M. Buxton, D. Buysse, D. F. Dinges, J. Gangwisch, M. A. Grandner, et al. Joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society on the recommended amount of sleep for a healthy adult: Methodology and discussion. *Sleep*, 38(8):1161–1183, 2015. https://doi.org/10.5665/sleep.4886.

[23] Dark Sky Company, The (LLC). Dark Sky, 2016. https://darksky.net/app.

[24] S. Diehl, F. Beck, and M. Burch. Uncovering strengths and weaknesses of radial visualizations - an empirical approach. *IEEE Transactions on Visualization and Computer Graphics (Proceedings of InfoVis)*, 16(6):935–942, 2010. https://doi.org/10.1109/TVCG.2010.209.

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

[26] G. M. Draper, Y. Livnat, and R. F. Riesenfeld. A survey of radial methods for information visualization. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 15(5):759–776, 2009. https://doi.org/10.1109/TVCG.2009.23.

[27] S. M. Drucker, D. Fisher, R. Sadana, J. Herron, and m. c. schraefel. TouchViz: A case study comparing two interfaces for data analytics on tablets. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 2301–2310, 2013. http://doi.acm.org/10.1145/2470654.2481318.

[28] J. Fuchs, F. Fischer, F. Mansmann, E. Bertini, and P. Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 3237–3246, 2013. http://doi.acm.org/10.1145/2470654.2466443.

[29] 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 (TVCG)*, 23(7):1863–1879, 2017. http://doi.org/10.1109/TVCG.2016.2549018.

[30] Garmin. Connect, 2018. https://connect.garmin.com/en-US.

[31] E. Hawkins. Climate Spirals, 2016. https://www.climate-lab-book.ac.uk/spirals.

[32] J. Heer and M. Bostock. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 203–212, 2010. http://doi.acm.org/10.1145/1753326.1753357.

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

[34] J. Jo, B. Lee, and J. Seo. WordlePlus: expanding Wordle's use through natural interaction and animation. *IEEE Computer Graphics and Applications (CG&A)*, 35(6):20–28, 2015. https://doi.org/10.1109/MCG.2015.113.

[35] Joyent, Inc. Node.js, 2018. https://nodejs.org.

[36] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 5092–5103, 2016. http://doi.acm.org/10.1145/2858036.2858558.

[37] T. Kekeritz. Weather Radials, 2014. http://www.weather-radials.com.

[38] Keming Labs. Weathertron, 2013. http://theweathertron.com.

[39] S. Y. Kim, Y. Jang, A. Mellema, D. S. Ebert, and T. Collinss. Visual analytics on mobile devices for emergency response. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pp. 35–42, 2007. http://doi.org/10.1109/VAST.2007.4388994.

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

[41] S. Kuijpers. Weather Eindhoven 2014, 2015. https://aka.ms/kuijpers-weather-radial.

[42] B. Lee, M. Brehmer, P. Isenberg, E. K. Choe, R. Langer, and R. Dachselt.

Data visualization on mobile devices. In *Extended Abstract Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 1–8, 2018. https://mobilevis.github.io.

[43] B. Lee, T. Dwyer, D. Baur, and X. G. Veira. Watches to augmented reality devices and gadgets for data-driven storytelling. In N. H. Riche, C. Hurter, N. Diakopoulos, and S. Carpendale, eds., *Data-Driven Storytelling*. A K Peters / CRC Press, 2018. https://aka.ms/dds_book.

[44] M. Lima. *The Book of Circles: Visualizing Spheres of Knowledge*. Princeton Architectural Press, 2017. https://aka.ms/lima2017.

[45] S. Lu. Annual Temp - New York 2015, 2017. https://aka.ms/lu-weather-radial.

[46] Microsoft, Inc. Azure Cloud Computing Platform & Services, 2018. https://azure.microsoft.com.

[47] T. Munzner. *Visualization Analysis and Design*. A K Peters Visualization Series, CRC press, 2014. http://www.cs.ubc.ca/~tmm/vadbook.

[48] OffCoast (LLC). Weather Line, 2018. http://weatherlineapp.com.

[49] R. Olson. What 12 months of record-setting temperatures looks like across the U.S. *FiveThirtyEight*, 2015. https://aka.ms/olson_fivethirtyeight.

[50] I. Ros and Bocoup. MobileVis, 2014. http://mobilev.is.

[51] S. Sadowski. Mobile InfoVis, 2015. http://mobileinfovis.com.

[52] J. Sauro and J. R. Lewis. Average task times in usability tests: What to report? In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, pp. 2347–2350, 2010. http://doi.acm.org/10.1145/1753326.1753679.

[53] B. Watson and V. Setlur. Emerging research in mobile visualization. In *Tutorial Proceedings of the ACM Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI)*, pp. 883–887, 2015. http://doi.acm.org/10.1145/2786567.2786571.

[54] H. Wickham and L. Stryjewski. 40 years of boxplots. Technical report, 2012. https://vita.had.co.nz/papers/boxplots.pdf.