CHAPTER 6

Evaluating Mobile Visualizations

Frank Bentley
Yahoo, USA

Eun Kyoung Choe
University of Maryland, College Park, USA

Lena Mamykina
Columbia University, USA

John Stasko
Georgia Institute of Technology, USA

Pourang Irani
University of Manitoba, Canada

CONTENTS

Evaluation is important for data visualization because it can help us to understand whether a system is achieving its intended goals and how well (or not) it is doing so. The evaluation of mobile visualizations builds upon research and knowledge gained from data visualization broadly, but it provides an additional set of challenges. There are many reasons to evaluate mobile visualizations with people, including potential

DOI: 10.1201/9781003090823-6

end users and experts. However, depending on the research questions that one wants to answer and the intended audience, different methods are needed. In the mobile information visualization domain, there is a broad continuum of research questions that can be answered by an evaluation study. Some goals include validating rapid perception of differences in data, while others focus on examining the long-term use of visualizations and whether seeing a visualization over months leads to their intended impact on users. Very different methods, time scales of research, and participant recruitment strategies are needed depending on the questions that one wants to answer. This chapter will explore the literature, discussing a variety of goals and evaluation approaches, highlighting best practices and making recommendations for future approaches to evaluating mobile visualizations.

6.1 INTRODUCTION

Evaluation is a vitally important yet extremely challenging aspect of data visualization [18, 70]. Visualization developers often want to compare their approach to that of others, or they may simply want to assess if their approach is achieving its intended goals. Visualization researchers have drawn from all that has been learned about evaluation from fields such as Human-Computer Interaction (HCI), Psychology, and others. Data visualization offers its own unique set of challenges, however, with many involving the difficulty of evaluating the utility of a visualization. Mobile visualization presents even more challenges including the purpose and context of use, diversity of potential users, and variety of display devices, to name just a few.

As discussed in Chapter 1, mobile data visualizations are developed for many different purposes, and researchers from many disciplines design and create mobile data visualizations. Whether they are interested in measuring human perception of differences in information presentation, trying to visualize a dataset in a new way to derive insights, or trying to create successful long term behavior change in a diverse set of users, researchers often have vastly different goals when setting out on a new visualization project.

Perhaps in part due to the different disciplines of researchers involved in creating mobile visualizations, a common set of best practices for their evaluation has not emerged. Researchers from the InfoVis, Mobile HCI, and Ubiquitous Computing communities, respectively, have traditionally used different sets of methods with different sets of research participants to evaluate their work. If one is only concerned about human perceptions of differences, long-term, longitudinal studies are not needed, and a short in-lab study can often suffice. However, behavior change is a long process that can take months (or even years) to measure, and therefore needs long-term studies with participants that carefully match the broader population to determine a system's effectiveness in real world conditions. Chapter 5 of this book further explores this continuum from perception to behavior change.

The domain of mobile devices brings new challenges to evaluating information visualizations. Mobile devices are more commonly used in the background throughout the day, in short bursts of activity, compared to longer and more purposeful interactions with desktop systems. Mobile devices also have a broad range of device characteristics, with differing screen sizes, resolutions, brightness levels, and other technical aspects that can alter a user's ability to perceive differences in a visualization. Ultra-small mobile devices, such as smartwatches, impose even greater constraints. The places where users choose to engage with mobile visualizations are also quite varied. Some might be used in bright sunlight while running, on a bumpy public bus, or while one's hands are full with other objects. These constraints mean that traditional, lab-based evaluation while seated at a desk with standard lighting conditions, are increasingly not appropriate methods to evaluate these systems that will be used in different situations and settings outside of an office.

Mobile devices are also ubiquitous in today's world. Therefore mobile visualizations often have an audience of the general public, over earlier work in the information visualization domain that was often focused on domain experts or other more highly educated individuals. When deploying systems to the general population, it is important to remember that a significant proportion of adults (studies have identified 41% of Americans as well as 44% of Germans [39]) have low graph literacy skills in understanding very simple bar and pie charts, and two-thirds of American adults do not have a bachelor's degree [17]. Therefore, with any visualization that one wishes the general public to use, it is important to recruit participants that match that broader population. Evaluations with college students or staff will lead to skewed results of use and understanding of data compared to a general population.

Furthermore, like the people using mobile devices, the devices themselves are quite diverse. Mobile displays vary in size, form factor, resolution, and input capabilities, as well as other characteristics. For example, many older mobile devices do not have the high-definition displays and the pixel scaling ratios of newer models, and thus cannot display visualizations at a comparable level of detail. Many people, especially those in developing countries, are using older phones as well as basic feature phones with smaller screens and limited touch capabilities. Addressing this diversity of devices in evaluations is important.

The remainder of this chapter will discuss existing best practices for evaluating visualizations, discuss our method for discovering and exploring different goals for evaluation, and then will explore best practices for each of these goals in turn. While there are obvious overlaps to the evaluation methods covered in Chapter 5, we cover goals and evaluation methods for mobile visualizations more broadly, going beyond what are necessary for glanceable visualizations. We will conclude with a discussion about the current state of mobile visualization evaluation and open areas of research in this domain.

6.2 BACKGROUND

The HCI community has made great strides toward effectively evaluating the usability of software systems. Techniques for determining whether a person can effectively use a piece of software are now commonplace and employed widely throughout the information technology community. Recent books about HCI lay out both methods and best practices for conducting such evaluations [60, 67].

The field of data visualization has leveraged much of that knowledge, but it also introduces some new challenges. One of the primary challenges in evaluating visualizations is the importance of determining a system's utility. While usability is often the evaluation focus of HCI, utility goes beyond it and toward the effective and beneficial application of the software for achieving greater goals. Specifically within the visualization domain, utility involves interpretation of visual representations of data for better understanding of the data and use in subsequent sensemaking and decision-making. Is a visualization helping viewers understand the depicted data better? Can viewers then make better decisions based on the improved data understanding? What kind of reaction, reflection, or memory does the visualization spur?

Plaisant [70] identifies the key challenges in evaluating visualizations and explains why this process is so difficult. Example issues that she calls out are the need for longer term assessments, for ways to judge whether a visualization can answer questions that a person does not know they have, and for methods to identify beneficial discoveries or general awareness facilitated by a visualization. She calls for more repositories of example data and tasks, as well as case studies and success stories.

Carpendale builds upon those thoughts and provides an in-depth tutorial of visualization evaluation methods and challenges [18]. She states that three key features are desirable to evaluation: generalizability, precision, and realism. She then describes a variety of quantitative and qualitative methods that visualization evaluators could utilize, explaining the methodological details, challenges, and threats to validity that each face. Ultimately, her goal is to provide guidance to others for choosing an evaluation approach.

When determining which evaluation method to use, fundamentally, one must first determine the ultimate goal of the evaluation. For example, is the goal to determine whether a person can accurately interpret what they see, or possibly it is to determine whether the visualization provides a greater understanding of the underlying data thus assisting subsequent decision-making. Depending on the objective of the evaluation, a variety of metrics may be appropriate to achieve that goal. Determining what to measure is one of the fundamental steps of any evaluation process. Of course, the next question is how to measure it. A wide variety of evaluation techniques have been proposed, with each providing specific benefits and challenges.

One of the most common evaluation approaches for visualizations is a controlled experiment. Study participants are given a series of benchmark tasks to perform while using a visualization. In some cases, multiple visualizations may be employed in a comparative study seeking to understand how they perform against each other. Potential metrics for this type of evaluation are the task success rate, time to completion, and number of errors made. These types of studies and their metrics are typically more quantitative in nature, and are viewed as being more objective. However, while these studies provide a large degree of *internal validity*, meaning that they can be replicated in the same exact lab conditions, they often lack *external validity*, meaning that they often cannot be replicated in real world situations with real devices and diverse people out in their natural contexts and activities.

Other forms of evaluation are much more observational and interpretive in nature. These types of studies usually seek *ecological validity* and thus emphasize observation of a visualization's use and application in realistic environments and scenarios. Frequently, these types of evaluations involve significant interaction with users of the visualization including interviews and capturing use over time. These types of evaluations are much more qualitative in nature and seek to accurately interpret and understand people's use of and opinions about the visualizations. However, these studies can also capture objective measures related to behavior change over time in order to augment the more qualitative data from interviews and diaries.

Shneiderman and Plaisant advocated for these more observational and interpretive methods, arguing that short-term lab-based studies simply will not be sufficient for evaluating visualizations. The researchers specifically articulate the need and motivations for longer-term studies of deployed visualization use [84]. They introduce the Multi-dimensional In-depth Long-term Case (MILC) methodology that is more like a case study, examining in-depth usage of a visualization by a few individuals. This type of evaluation has gained increasing traction within the visualization community.

Another form of evaluation does not involve studies with sets of target audience, but instead employs expert reviews of a software system. The Insight-based evaluation methodology is a hybrid form of a controlled experiment and an expert review. It attempts to quantify the number of insights generated in the use of a visualization [81]. Evaluation participants, who must be at least knowledgeable in the domain of the data being presented, use different visualizations while the evaluators count and measure information such as the time to reach the first insight, count of insights, and domain value of the insights. Although it is conceptually appealing, this technique is quite methodologically challenging to implement.

More recently, Wall et al. proposed the ICE-T evaluation methodology that provides a heuristic-based discount evaluation approach [88], a form of expert review. The technique employs a hierarchy of evaluation heuristics under four primary components: insight, confidence, essence, and time. Evaluators who bring knowledge of data visualization rate a system on each of the heuristics, ultimately producing a score for the visualization that estimates its value and utility for helping to understand the underlying data. The technique is designed to be fast and easy to deploy, not requiring interactions with potential users.

More recently, researchers have begun to look beyond performance-based metrics such as a visualization's usability and utility [80, 89]. Alternative metrics such as the memorability of a visualization or the engagement, and the enjoyment it provides, also may be important.

Lam et al. conducted a meta-study of information visualization evaluations, drawing from 361 academic papers discussing some type of visualization evaluation [58]. The researchers identified seven primary evaluation scenarios under two high-level goals: understanding data analysis and understanding visualizations. For each of the seven categories, they then describe its goals and outputs, evaluation questions, methods, and list examples. Isenberg et al. subsequently expanded this review to other subareas of visualization [47]. The authors reported that measuring visualization system's resulting images and its algorithmic performance has been the most predominant evaluation methods, and only recently, there has been a steady increase in evaluation methods that include participants.

Moving beyond work in Data Visualization, the Ubiquitous Computing and Mobile HCI communities also have developed a large set of methods for use in evaluating the use of technology in the wild. The goals of many of these studies have been to understand how technology is adopted into everyday life and how people alter their behavior when given a new system. Many of these studies therefore employ short and long-term field study methods, where participants interact with a system throughout their own daily life as they see fit, and the evaluation is meant to measure their understanding of the data collected, specific changes to behavior that have been made, or to understand if enough data can be collected at high quality to make a system feasible for use in the world.

Diary Studies [68] are frequently used to understand specific actions that people take in a system over time when interacting with a system naturally outside of controlled lab conditions. Participants can leave voicemail entries, complete paper worksheets, or online surveys to leave details of specific interactions that they had with the system close to the time of usage, while memories are still fresh. Diaries are often combined with Experience Sampling Method (ESM) [3, 30], where an application prompts a participant for feedback about their experiences at random times throughout the day or after specific events take place (e.g., after logging food in a diet tracking application). Both of these methods are particularly important for studying the use of mobile systems, as interactions take place in the world over time and are not directly observable by researchers. Gathering data about a person's experience as close as possible to the interaction helps to improve the accuracy of the data collected as experiences are recent and not recalled weeks or months later in an interview. A thorough review of both of these methods can be found in Consolvo et al.'s Mobile User Research book [26].

Other field study methods can involve Contextual Inquiry [46], where researchers travel to locations where participants typically engage in an activity and study existing practices or the use of a new system within that context. This method can help researchers to understand environmental variables (e.g., noise, light, distractions, network conditions) that can impact use of a system in real world contexts.

Finally, often researchers need to conduct a formal experimental study to validate that a design meets a specific objective. Whether this is a classic A/B test, first explored by Google in 2000, where a random sample of users get one experience and the rest get a control, or a more complex experimental design [44], experiments can prove that a given experience can change a user's behavior in a statistically measurable way. Often, this is required for field studies of behavior change systems in order to prove that they perform better than existing approaches or interview-based methods without a system in place.

Overall, these varied methods have been applied to a wide range of systems by researchers working in several adjacent disciplines. However, what was lacking in the community was a more systematic investigation of how these methods could be used in the evaluation of mobile visualizations in particular. The remainder of the chapter will explore this topic.

6.3 METHOD

We conducted a literature review through searching on Google Scholar and the ACM Digital library to identify papers that include a visualization on a mobile device and were evaluated directly with end users in some way. This search included both directed search for papers that we were familiar with as long-time researchers in the field, branching off from those papers based on citations, and more general searches of the databases for keywords related to mobile visualizations. We also looked through proceedings of recent Ubicomp/IMWUT, Mobile HCI, CHI, and InfoVis conferences.

In total, we reviewed 31 papers that met our criteria and represented a wide range of systems from each of the research communities that create and evaluate mobile visualizations. For each paper that we found, we identified the evaluation method used as well as the primary goal of the evaluation (e.g., perception, usability, behavior change). These codes for the primary goal evolved as we examined additional papers until we arrived at the set of six that are explored in the next section.

We further examined papers in each of the goal areas to identify common methods and best practices for research that can answer a variety of types of research questions. For example, very different methods are required if one wants to see if participants can perceive a difference between two bars compared to one that is trying to quantify behavior changes that participants make over a 90-day period as a result of having a particular visualization in their lives.

The following sections will introduce the framework of six evaluation goals for studying mobile visualization systems and then will explore each goal in additional detail, covering the methods and typical study designs used when faced with each evaluation goal.

6.4 FRAMEWORK

After reviewing the literature using the method above, we settled on six goals for conducting evaluations of mobile visualizations. These goals, as shown in Table 6.1, form a rough hierarchy, whereby questions at the top of the list should be explored before answering more involved questions below.

Perception: At its simplest, the most basic goal of evaluating a visualization is to see if a participant can perceive it. Can they recognize what is shown and differences between data representations? For example, one might want to know if a participant can tell that one slice of a circular graph is larger than another.

Usability: The next goal is to see if a participant can interact with the visualization and if the visualization is usable. For example, can a participant navigate to the correct data, filter it in meaningful ways, and understand the various controls and UI elements of the visualization? It is important to understand if a visualization is usable before going through the effort of running a feasibility study, as quick design iterations can improve the interaction and increase the likelihood that users will stick with the system long enough to capture the data needed to support long term goals.

Feasibility: Mobile visualizations often are powered with data that is collected out in the world, often on mobile devices themselves. The next goal is to evaluate

TABLE 6.1 The six goals for evaluating mobile visualizations with example projects highlighting the methods used to evaluate systems for each of the goals.

Goals of Evaluation	**Examples**
Perception (Can the user recognize what is shown in the visualization?)	SmartWatch Vis [9], Residential Energy Use [74], Ambient Phone [82]
Usability (Can the user interact with the visualization and understand its controls?)	Google Maps [73], Minimap [76], Vis-Tiles [59], User Interactions with Scatterplots [15]
Feasibility (Can the data be collected/cleaned to power this visualization?)	OmniTrack [54], Mobile Health Mashups Pilot [85], Patterns of Everyday Life [29]
Understanding (Can a user understand the state of the data and what it represents?)	Whereabouts Clock [14], Motion Presence [8], GSparks [64], inAir [53], Ranges over Time [12], Animation vs. Small Multiple [11], GlucOracle [31], When(ish) is My Bus [52]
Reflection and Insights (Can a user learn new things from the visualization?)	TummyTrials [50], Lullaby [51], SleepTight [22], Tangere [77], Pass the Ball [75], VisualizedSelf [24], Orchard [33]
Behavior Change (Can the user change their behavior given accurate interpretations of the visualization?)	Health Mashups [7], UbiFit [28], Ubi-Green [38], Fish'n'Steps [61], How to Nudge InSitu [48], BeWell [16], Glanceable Feedback for Physical Activity Tracker [41], WaterJewel [37], Persuasive Performance Feedback [23]

if a visualization is feasible. Can the appropriate amount and accuracy of data be collected to lead to meaningful experiences? These types of evaluations might have a participant collect data for several weeks, and then see what could be populated into their visualization to see if real-world data collection can power the design.

Understanding: The final three goals of mobile evaluation involve participants being able to take meaning away from the visualization. The most basic of these goals involves understanding the real-world meaning behind what they are seeing. Where the basic goal of perception is typically agnostic of the data (e.g., bar 1 is longer than bar 2), this goal involves understanding what the data actually means (e.g., you walked a little more on Tuesday compared to Wednesday).

Reflection and Insights: Moving beyond understanding comes reflection and data insights. Here, the goal is to see if participants can go beyond the data itself to uncover something more complex about their lives or the underlying data. For example, a participant might see that every Sunday he walks less than other days of the week, or that a particular variable is trending up or down over time.

Behavior Change: Finally, the most complex goal is understanding if participants are changing their behavior based on the visualization. In these evaluations, participants often live with a design in their life for weeks or months, and the goal is to see if the visualization leads them to change behaviors—to walk more, drink less, find a faster route through traffic, etc.

These six goals are not always explored in isolation. As we will discuss with the examples in the next section, often a project will span several goals—for example, exploring perception and understanding together, or data insights and behavior change together. However they do build on each other. If a given system does not meet the goals on the top of the list, then it will be very difficult to meet the more complex goals below. For example, if a person cannot perceive a difference or understand the data, it is quite unlikely that they can produce accurate data insights or change their behavior based on a trustworthy interpretation of the data.

6.5 EVALUATION METHODS

We will now move through Table 6.1, exploring example research projects as case studies under each evaluation goal as well as best practices for choosing methods, research participants, durations of studies, and evaluation metrics for each category.

6.5.1 Perception

One of the more fundamental criteria of a good visual display is whether it is perceptually legible. Activating the appropriate perceptual mechanisms is vital to identifying key aspects of a graph or a visualization. As such, designers can guide their choices based on low-level visual processing of the information display [90]. The perceptibility of visual information on mobile, often smaller displays, is distinct from our understanding of guidelines for presenting graphs on paper, or even on a fixed desktop screen. Often, visual content is consumed while moving or on-the-go and this alone can generate significant variability in terms of screen glare and color disturbance, as much data has to be compressed on a smaller display. Further, these visualizations frequently accommodate a primary task the user may be engaged with, such as jogging, and therefore users can possibly be less attentive to the display, making the need for fast and accurate perception of the data critical. The smaller display or displays of varying form-factors and resolutions further exacerbate the designer's task for identifying the optimal rules under which to display data. Under these constraints it is evident that a first stage that involves assessing the perceptual limits or perceptibility of a design is necessary for suitable mobile information displays.

Perceptual assessments were initially popular when mobile visual displays were primarily geared at showing spatial information, such as points-of-interest on a map [21]. For such a problem, choosing any non-trivial display can often result in portraying far too much in a small display region. A common approach involves filtering the data and navigating to elements of interest. However, this does not eliminate the necessity to visually parse through content, as it is cognitively demanding to pan and zoom through a display [15], but further also forces users to lose the global context when inspecting details. The small display in such types of datasets also results in placing content outside the display's viewport, also commonly referred to as off-screen content. Techniques such as Halos [4] resolve this concern by placing concentric circles centered at the off-screen points-of-interest. Furthermore, the Wedge [42] relies on the Gestalt principle of amodal completion [34] by providing a small section of the visible arc, along with two 'legs' that would form a triangle, with the tip placed at the off-screen object location. In validating their visualizations, the researchers resorted to asking participants to provide their best estimate of the off-screen target locations, but on an emulated mobile device presented on a desktop. While mobile emulation can assess the perceptual limits of such techniques, it lacks the degree of ecological validity needed to fully capture visualization when viewed on an actual mobile device. Furthermore, such perceptual tasks do not capture on-the-go contexts, as the targets were placed in static positions. To offer a more ecologically valid task, Gustafson and Irani [43] made the points-of-interest 'disappear' as would be common when the user would move through an environment. Their evaluation was also carried out on a mobile emulation, lacking the rigour needed to fully gauge the ability to quickly glance and perceive the visual changes on the screen.

While such earlier perceptual evaluations were geared at understanding the effectiveness of the visualizations, current interest for displaying content on mobile devices has predominantly focused toward the evaluation of data charts [10]. Graphs, including line graphs, bar charts and even pie or donut graphs, have gained recent popularity for their ability to scale in size while still preserving a general sense of the content. These are often displayed on very small smartwatch displays, for the general consumer interested in taking a peek at their heart rate or other measurable biometrics, often produced by the worn device. They are not spatially constrained, as with points-of-interest on a map, but yet have to portray as accurately as possible the underlying dataset. Producing generic guidelines for displaying content suitably on such small devices is challenging, as smartwatches are extremely limited in terms of pixel resolution. Yet, these visual displays need to be sufficiently glanceable (to enable quick perception of the key graph content). Furthermore, smartwatches being either circular or rectilinear add additional perceptual constraints on how best to present the graph [64].

Inspiration can be drawn from a wide number of visualization techniques that have been designed and proposed and that could be suitable for small screens. For example, that specifically compress data pixels along a given axis (usually the y-axis) have be adapted to displaying charts on small screens.

Drawing inspiration from techniques such as HorizonGraphs [79] and SparkLines [87], Neshati et al. introduced G-Sparks [64] that compresses the visual pixels to present as much as possible on a small screen. Their evaluations were carried out directly on a smartwatch display, thus offering a more ecological valid baseline for future designs. Similarly, Blascheck et al. performed a perceptual study that asked participants to compare data points from three chart types, on a smartwatch display, while seated a fixed distance away from the device [9]. The researchers sought to find out how quickly people can perform a very basic task—to ultimately understand whether certain charts might be useful in situations where people can only quickly glance at a watch (e.g., while doing sports, driving, etc.). Their results suggested that Bar and Donut charts performed similarly while Radial Bar charts performed the worst. They provide guidelines on the glanceability of such graphs for small displays.

Pushing the tradition of graphical perception experiments further, Brehmer et al. conducted a series of crowdsourced experiments on participants' mobile devices [12, 11]. They demonstrated that it is feasible to directly deploy experimental software on participants' mobile devices to run controlled experiments and to collect measures such as task completion time, error, and subjective feedback. To identify effective ways to present ranges over time on small displays, the researchers compared two visual encodings for time-oriented data—linear layout vs. radial layout, while also varying the data granularity, data sources, and task types [12]. In a later study, the researchers conducted a comparative evaluation of animation and small multiples for trend visualization on mobile devices [11]. In both studies, the research team leveraged Amazon Mechanical Turk (AMT), a popular crowdsource platform, to recruit about 100 participants. To ensure instruction compliance and data quality, various techniques (e.g., minimum instruction reading time, interruption monitoring) were embedded in the experimental platform. Participants recruited via AMT were in a range of physical settings using a diversity of devices, ideally being more reflective of when and where they would use their mobile device, with no physical presence of an observer watching them. Such a method comes with a loss in consistency, and like college student participants, crowdworkers constitute a specific demographic that may not be representative of the intended target population for the visualization. However, with quality control measures in place, mobile visualization evaluation applications deployed via crowdsourcing platforms have the edge in its capability to recruit many participants within a short time period.

Previously proposed studies evaluating the visual efficacy of common graphs on small displays often limit the assessment, for practical reasons, to only a few conditions. Namely, tasks given to participants require a comparison among different points, such as identifying the minimum or maximum within the graph, detecting the slope direction, or differences among points [9]. Such tasks have been commonly used for assessing the legibility of data graphs on typical displays. Furthermore, current devices are equipped with legacy visual designs, such as bar or line graphs. However, it is unclear whether these are also the best suited for tasks needed by participants wearing small displays. Amini et al. [1] sought to identify what may be common tasks participants could benefit from in mobile conditions, and how to support these visually. The authors collected such tasks from those already equipped with a smartwatch and asked designers to produce suitable visual presentations. Very few among these were considered standard among the visualizations used on such small displays. This exercise thus leads to rethinking our approach to evaluating visual designs, by first developing an initial understanding of commonly needed tasks, and then adapting our presentation to these.

6.5.2 Usability

As the mobile phone has become the de facto platform for people to access information, conducting usability evaluations in a mobile environment has become common as well. Usability evaluation is a critical step to ensure that the user interface designed for a mobile environment functions as we expect and meets the requirements of the end users. Tasks for mobile usability studies deal with *why* and *how* people interact with information on the screen, as well as *what* people enter [13]. In a traditional usability study, researchers measure how effectively or efficiently a system supports people's tasks, typically in a controlled lab setting [32]. Usability is measured through success rate (whether participants can perform the task at all), task completion time, error rate, and participant's subjective satisfaction [65]. Collecting such measures will help companies, researchers, or developers assess their original design in comparison to a redesign or a competitor' product, and help improve the design of the interface iteratively.

In comparison to evaluating desktop-based applications, usability testing for mobile is different in three aspects. First, special equipment (e.g., document camera, webcam) and software may be required to record the mobile screen, which would allow real-time projection and monitoring of the mobile screen. Sometimes, mobile (e.g., Eyezag [36]) and head-mounted (e.g., Tobii Pro [72]) eye-tracking devices are used to capture people's natural viewing behavior or "areas of interest." Incorporating eye-tracking in a usability evaluation, researchers can capture a comprehensive picture of people's experience with a mobile user interface, as in the case of Cheng et al.'s usability study, which was aimed to evaluate people's mobile browsing and searching behaviors on a mobile device [20].

Second, as mobile devices are used across a broad range of places and contexts (e.g., in the car, on the street, standing or walking), creating a realistic mobility testing environment with high ecological validity may be necessary. To create such usability testing environments, researchers have created simulated walking conditions using a treadmill (e.g., [62, 2]) or conducted studies on a closed course (e.g., hallway) (e.g., [63]). Notably, Kane et al. created a "Walking user interfaces" (WUIs) that adapt their layout (e.g., button sizes) when the participant is moving [49]). WUIs were evaluated in an outdoor public open plaza where other people besides the participant were frequently standing, walking, and interacting with one another. Hincapié-Ramos and Irani also designed WUIs, which provide safety visual cues of obstacles to prevent collision for those who use mobile phones while walking [45]. To evaluate the system in a realistic setting, participants were asked to walk down the university cafeteria where an "actor" was present to provoke potential collisions (e.g., cut the participants' path orthogonally) [45].

A third aspect of interest in mobile usability studies is the additional focus of the assessment—the different interaction repertoire available on or with mobile devices that are not available in desktop solutions. Such interactions include cross-device interaction techniques leveraging spatially-aware mobile devices to counteract the limited screen space of single mobile device [59]. In Langner et al.'s work, a preliminary usability testing was performed to get quick feedback on the key interaction concepts and workflow before they built the final prototype [59].

The number of participants recruited in these mobile usability studies are similar to that of typical usability study counterparts, ranging from 10 to 30 participants. Sometimes, mobile usability study can be done completely remotely on the participant's phone. In such a case, detailed logging or installing software (e.g., TeamViewer) to project participants' mobile screen is necessary to capture their behaviors.

Evaluating usability will always be an important step toward designing novel visualization systems. Using creative methods like the ones we mentioned above will particularly be helpful in situating and capturing people's realistic mobile experience with smartphones and smartwatches, providing us with useful insights that might have not been possible from employing traditional usability testing methods.

6.5.3 Feasibility

When developing a new visualization technique, or applying an established one to a new domain or means of data collection, evaluating if you can feasibly collect the amount and quality of data that you need to power the visualization is often necessary. Mobile visualizations add to this need since the data that powers the visualization is often collected from the phone or physical environment over time and can often be spotty or inaccurate due to the nature of consumer-grade mobile sensors. There are several aspects that researchers need to address when checking for the feasibility of a mobile visualization. These include compliance rates, retrieving data from sensors in real world situations, and data quality.

If a system involves participants needing to manually log or upload data, understanding the compliance rate is critical to understand if it will provide enough data, and that the data is not biased to particular days or times of day. For example, if participants only provide data in the evenings, it might not represent their behaviors throughout the day. Or if participants do not engage in providing data on weekends, key aspects of their lives may be left out. It is also important to ensure that you can collect enough data to power your visualization. If you need 100 data points, how long will it take participants to provide that many? What attrition might occur due to fatigue or lack of benefits until enough data is provided?

When evaluating a system that collects sensed data (e.g., heart rate, location), understanding the error rates, sampling frequencies, and battery impact to the device might be necessary in order to understand if a given visualization is feasible. Mobile sensors, such as location, often have error margins associated with them. Location in particular can have several hundreds of meters of error when indoors or when using network-based location technologies when saving power. Using GPS frequently could drain a participant's battery. Other sensors, such as pedometers might have greater than 20% error [86], or participants might not wear them every day or for every activity [27]. Feasibility evaluations can help researchers to understand the data that particular sensors can provide in real-world situations.

Since most visualizations require a certain amount or quality of data to function, it is critical to understand if your system will provide appropriate data before continuing on with a longer and likely more expensive field evaluation, for example those in the behavior change section below. It is also critical to be able to show or mimic realistic data when performing studies of visualization understanding (in the following subsection). Presenting participants with overly idealistic data will not lead to valid conclusions about how your system will function in the wild with real data.

Several methods can be used to check the feasibility of a visualization, mostly depending on where the data is coming from. If the data is coming from end users, these studies typically include pilot field studies where a small group of participants, typically around 10 to 15, interact with your system for a few weeks. It is important that these participants are representative of your target user base (e.g., in terms of education, data and smartphone literacy, age, physical abilities). If you are aiming for mass-market adoption, ensuring that 2/3 of your participants do not have a college degree is important, as well as age, gender, and racial balancing to match the broader population as closely as you can, as any of these demographic or cultural factors might impact compliance rates and overall data quality.

An example of a feasibility study is the pilot for the Health Mashups project [85]. In this study, ten participants used a system to collect data about their wellbeing from a variety of sensors for a 4-week baseline and 4-week study period. The goal was to understand if participants would provide enough data for statistically significant correlations to be made between streams of data, for example being able to tell a participant that they walk less on hot days or that they gain weight on Sundays. A reasonable amount of data was received from automatic tracking sources such as the Fitbit for step count and the scale for weight, however the researchers discovered that participants were not providing enough samples of manually logged data such as food intake or mood. As a result, a new mechanism was designed using silent push notifications for reminding participants to log these [6], which led to a significant increase in logging in the full study [7]. Without this feasibility study and resulting design change, time and money would have been wasted in running a larger and longer behavior change focused study, as enough data would not have been provided for these key aspects of wellbeing. This highlights the importance of checking more basic goals in the framework before moving on to the more complex goals.

In Omnitrack [54], researchers explored how a flexible, personalized tracking platform can support people's diverse tracking goals. Using OmniTrack, participants created their own tracking systems for a wide variety of goals (from tracking beer drinking to quality of sleep) and used the trackers during the two-week field study. The field study showed the feasibility of a personalized tracking approach in terms of compliance and the amount of data collected. Correia et al. [29], in another example of an evaluation to measure feasibility, had participants use a location-capturing app for four days, and then looked to see if the data was sufficient to power their visualization showing patterns of everyday life.

Whitlock et al. [91] illustrated another way to assess a visualization's feasibility, among other measures. The researchers developed a visual analytics system, deployed on mobile devices, to assist with data gathering in the field in domains such as earth science and emergency response. They used a form of expert review to evaluate their prototype. More specifically, they demonstrated the system to ten field analysts and had each work through example tasks and scenarios. Finally, the researchers conducted in-depth interviews with each analyst to gauge their response to and opinions about the system.

These examples highlight the need to ensure that mobile systems can capture enough data with high enough accuracy in order to power a particular mobile visualization. Ensuring the users or the device can, in typical use, provide meaningful amounts of data is critical before moving on to later, more complex goals. Most importantly, this step gives researchers an idea of the amount of data needed to power the system, such that later evaluations can be designed around typical data that the system would provide, and not idealistic designs that might be more precise or contain more data than the system would produce when running in the wild with a diverse population of users in everyday conditions.

6.5.4 Understanding

After knowing that gathering the data for a given visualization is feasible, it is then necessary to know if users can understand the data presented within that visualization. Beyond just perceiving differences in bars or data (e.g., the bar on the right is smaller than the one on the left), these evaluations seek to understand if typical users can understand what the data actually means (e.g., I walked fewer steps today compared to yesterday) for realistic data sets with data density similar to what was shown to be feasible.

Often, studies with these goals will deploy a combination of in-lab or online evaluation and short-term field studies with the goal of understanding how participants perceive the data that is shown. Can they answer specific questions about the data itself after seeing the visualization? Do they understand the meaning of the various bars or charts present in the visualization? Again, in these types of studies, it is important that the participants who participate are representative of your target user base. If you intend to release your system to the general public, that means that two-thirds of your participants should not have college degrees and they should span a wide range of ages and ethnic backgrounds. Since a significant proportion of adults cannot interpret a standard time-series graph [39], it is critical to get a sample of participants that matches the broader population for solutions intended to be used outside of elite circles of trained and data-literate professionals. As Peck et al. [69] show that including rural participants and those of varying educational and political backgrounds can yield diverse reactions to understanding and engaging with visualizations.

Simple studies can involve in-lab study or online studies that ask participants to onboard to a system via a natural means (e.g., downloading and installing an app, signing up on a website) and then looking at a visualization in the context of that experience and explaining what it is showing. Thereafter, specific pointed questions might be asked to ensure that participants understand what they are seeing. These types of studies can often be conducted with a few dozen diverse participants from outside of an academic environment.

On the other end of study complexity, systems can be deployed in the lives of users, and interviews can be conducted after a participant engages with a system for several days or weeks. This might be necessary if the data that is collected needs to be related to a user's life, such as with health data such as step counts. Participants should be able to understand what a visualization says about their own life and activities, which might be easier than looking at abstract data in a lab setting that is not related to their own experience.

Examples of these types of studies include "When(ish) is my bus" [52], where participants saw a probabilistic distribution of bus arrival times. In this study, participants were asked real-world questions such as if they had enough time to get a coffee before the bus was likely to arrive, rating the chance that the bus would show up within the next 10 minutes. The researchers had 500 participants view this visualization online and respond to these questions about data understanding. Consolvo et al. [28] explored the understandability of their Ubifit Garden visualization, prior to a larger field study focused on behavior change, through an online survey with 75 diverse participants.

Another example of this style of testing is the evaluation conducted for the Tangere visualization system [78] running on iPad tablets. Tangere supported up to three simultaneous views of scatterplot, barchart, and line graph visualizations. In this evaluation [77], participants received a set of 12 tasks (questions to answer) about a beverage sales data set. What was unique about this evaluation was that it was a "discoverability-focused" study. The researchers argued that for mobile interfaces running on a tablet, most people would not spend time working through tutorials and user manuals. Instead, they would simply open the interface and start working. Thus, the researchers followed a protocol just like that—participants only had a very short exposure to the system interface before attempting the tasks. Ultimately, the study provided significant qualitative data about the discoverability, learnability, and usability of interface gestures in terms of supporting how data was understood, as well as insights about the design of specific interactive operations.

The evaluation of Tangere utilized minimal onboarding, instead focusing on discoverability. Brehmer et al. have conducted studies in which they introduced participants to a visualization with incremental tutorials and asked screening questions to assess whether the participants understood how to read the visualization before proceeding with the experimental tasks [12, 11]. Onboarding user interfaces and tutorial tasks are common in (mobile) games, and it is worth considering how visualization designers evaluate comparable aspects of mobile apps that feature visualization.

Studies focused on measuring user understanding can also include exploring how users understand limitations or missing data for responsive visualizations. As discussed in Chapter 2, the responsiveness of a mobile visualization is its ability to adapt to different screen sizes and configurations. Responsive visualization design strategies include reducing the amount of data to show, aggregation, and changing the encoding. The designers of a mobile visualization may wonder whether a user on a mobile interface that omits key data points comes to the same understanding of the data as a user who views that same dataset on a desktop interface. But how should one evaluate the responsiveness of visualization? An evaluation strategy could involve quantifying the difference in the amount of information conveyed by visualization on a larger display and the amount of information conveyed on a mobile display. An example of this type of study is Schwab et al.'s evaluation of pan and zoom timelines and sliders in which they experimentally compare interaction alternatives on both desktop and mobile platforms [83].

The trickiest thing to get right with these types of studies is the research protocol. Making sure that questions are not leading, represent real world scenarios, and do not give away answers to future questions is key. The When(ish) is my bus study was a good example of this, asking real world questions about having enough time to get a coffee instead of priming users to understand that the visualization was showing variance in arrival time or giving hints about what either axis represented. Thinking about what you want the user to be able to accomplish with the visualization and ask about those real world goals instead of focusing on specifics of the visualization itself can ensure that you are measuring true user understanding of the visualization and not their ability to follow researcher directions.

6.5.5 Reflection and Insights

The utility of visualizations is greater than a mere conveyance of data. As shown in Table 6.1, many of the visualizations are designed to aid in awareness, decision making, or behavior change for the general audiences. Just because people can "comprehend" visualization does not necessarily mean that they can act on it. Visualizations may facilitate such transition from comprehension to action by supporting people to reflect on the data and generate insights from it. As such, we can assess visualizations' effectiveness by measuring its ability to support the process of self-reflection [24] and insight generation [66].

Visualization systems' capability to facilitate self-reflection has been studied in the personal data visualization contexts, in which people interact with the data about themselves, often with mobile and wearable devices. In prior works, researchers conceptualized different types of self-reflection (e.g., reflection-in-action vs. reflection-on-action [71]) and employed a variety of approaches to facilitate self-reflection, the outcome of which is personal insights. For example, Choe et al. designed a visual data exploration platform called Visualized Self that integrates personal data from multiple sources and promotes self-reflection [24]. Through an in-lab, think-aloud study, the researchers identified common personal insight types (e.g., external context, confirmation, contradiction, comparison, value judgment) people generated from interacting with Visualized Self. In Karkar et al.'s work, the authors proposed a framework for self-experimentation and designed a mobile application that provides personal insights on triggers for irritable bowel syndrome (IBS) [50]. In the same domain of providing individualized insights for IBS patients, Chung et al. explored the use of lightweight food diaries, which helped both patients and experts develop individualized, actionable plans and strategies [25]. In all of the works above, people's active involvement in data collection and exploration were essential in promoting self-reflection, which led to personalized insights.

In the visualization community, North suggested "insight-based evaluation" as a practical alternative to traditional evaluation methods (e.g., usability test, controlled experiment) [66]. Insight-based evaluation aims to directly capture insight while preserving the positive aspects of the traditional evaluation methods: participants are given initial questions, but they are encouraged to freely explore the data going beyond the initial question until they feel that they have learned everything from a given dataset. During this task, participants—typically domain experts—verbalize their findings, which are later segmented as an insight occurrence and coded based on a variety of metrics, such as insight category, complexity, time to generate, and errors. Insight-based evaluation is in general more time consuming than traditional experiments for both evaluators and participants. However, it can provide rich understanding on visualization's effectiveness in helping people generate insights.

While there is no seminal example of insight-based evaluation of mobile visualization, we believe such a method can be particularly helpful in evaluating domain-specific mobile information visualization targeting a specific group (e.g., patients, domain experts), or when there is no reasonable and fair counterpart to compare against the mobile device. For example, in evaluating personal informatics applications, we can tag insight-based evaluation onto a field deployment study, especially during the exit interview by asking what new insights the tool has enabled people to generate. In this way, participants can share new insights they gained throughout interacting with the new tool over a prolonged period, which can subsequently be analyzed based on the meaningful criteria for the domain.

6.5.6 Behavior Change

As with many interactive computing systems, the ultimate goal of many mobile visualizations is to assist people in achieving some goals through action. These goals and actions can vary greatly depending on the focus and purpose of the visualization. As a result, studies that focus on evaluating the impact of mobile visualizations on people's actions vary in their characteristics. There are a number of considerations that often influence the design of evaluation studies that focus on influencing people's action.

Depending on the type of action the visualizations aim to support, the timeframe of an evaluation study may vary from short-term to longitudinal. For example, for visualizations that aim to influence purchasing behaviors that are often made in the moment, evaluation studies could examine the immediate impact of visualization on individuals' purchasing choice. Following this idea, Kalnikaite et al. evaluated the impact of lambent devices that deliver salient information while shopping on individuals' purchases [48]; while this study included two separate shopping trips, the researchers captured each purchasing decision made while using the visualization. However, studies that focus on changing lifestyle behaviors and habits require considerably longer timeframes, on the order of months, as lifestyle behaviors change slowly. This applies, for example, to visualizations that target health behaviors, such as sleep, exercise, or diet. The evaluation study by King et al. [55] examined the impact of mobile visualizations on individuals exercise and sleep pattern in a study that included a 3-week run-in period and 8-week evaluation period.

Another important consideration is the setting in which the study is conducted. To a large degree, studies that aim to examine the impact of mobile visualizations on actions, choices, and behaviors are conducted in naturalistic settings and in the context where such decisions are typically made [19]. This approach allows researchers to minimize the impact of laboratory settings and argue for ecological validity of their findings. While, arguably, it is possible to evaluate such impact in controlled settings, the ecological validity of such studies will be limited.

In some cases, achievement of goals can be measured with standardized and validated questionnaires, which could be deployed at the baseline and upon completion of the study. For example, if users' goals include changes in psycho-social characteristics, such as self-efficacy, or changes in opinions and perceptions, such as satisfaction, there exist multiple validated questionnaires for measuring these outcomes in a reliable and consistent manner. However, oftentimes, for studies that target changes in users' behaviors, it becomes necessary to capture these target behaviors and objectively assess individuals' achievement of their goals. Gouveia et al. [40] evaluated the impact of a glanceable display of physical activity on individuals' achievement of personal activity goals as measured by the number of steps captured in a 28-day study.

In cases when researchers wish to evaluate the ability of visualization to influence users' behaviors, it is important to use a study design that allows one to differentiate between the impact of the visualization from other factors, for example from a simple influence of participating in a research study. In these cases, it often becomes necessary to introduce a control group, in which participants are exposed to all the other aspects of being in the study, but are not exposed to the mobile visualization under evaluation. Furthermore, randomizing participants into a control group and an experimental group helps to minimize potential differences between groups due to other factors, such as age, gender, or experience with technology. These study designs are typically called between-subject studies. An alternative approach allows researchers to use a single group, but randomize participants into receiving or not receiving an intervention, at different times, and comparing participants behaviors with and without the intervention [57]. These designs are often called micro-randomization. While these designs are not common for evaluating mobile visualizations, they are often used to evaluate other types of mobile technologies for behavior change. While not all evaluation studies rely on a comparison with a control group, simpler one-group pre-post study designs provide less reliable results.

Finally, another critical consideration is the strength of evidence needed to support researchers' claims, which determines the scale of the study and the rigor with which it was conducted. If the main goal of the researchers is to assert whether the visualization can have the intended impact in a limited setting and under a certain set of conditions, they may be able to support these claims with a small study. The majority of studies of mobile visualizations we reviewed would fall into this category. However, stronger claims that may suggest that such visualizations are ready for wide-scale dissemination and adoption require larger-scale studies with tests of statistical significance of their results, and analysis of statistical power.

Of all the evaluation studies reviewed in this chapter, these studies are typically the most challenging and expensive to conduct. They require full implementation of the mobile visualization in a robust enough way to be used by multiple individuals over extended periods of time. Furthermore, they may require additional development costs if researchers wish to isolate the impact of the visualization itself from the impact of the application that hosts the visualization. In this case, the researchers may need two versions of the application, one with the visualization and one without. However, these types of evaluation studies also provide the most direct evidence as to the ability of mobile visualizations to achieve their intended purpose and help their users achieve their goals. At the same time, we urge researchers and practitioners to not jump into a study to assess a system's impact on behavior change. As we discussed in Section 6.4, even though the system's ultimate goal is to demonstrate behavior change in the long term, if a given system does not meet the basic goals, such as perception, usability, and feasibility, it will be very difficult to meet the more complex goals such as behavior change. Klasnja et al. also discuss the importance of separating these different goals for different stages of design and indicate that HCI contributions should focus on efficacy evaluations that are tailored to a specific behavior change technique because results from such studies can provide in-depth insights on *why* and **how** a certain design decision works [56].

6.6 DISCUSSION

In this chapter, we have reviewed a rich spectrum of goals for evaluating mobile visualizations that come from multiple academic disciplines. Each of these approaches can solve specific needs and research questions that projects may have in various states of their development. As more basic goals need to be met before more complex goals can be tackled, we hope that this review can encourage researchers to adapt methods from adjacent disciplines to move evaluation beyond one specific goal to explore more complex goals along the spectrum.

We have explored the new constraints that mobile brings to the evaluation of visualizations, including issues with data capture in the wild and the wide variety of different handsets that users might have (and their differences in screen resolution, size, brightness, default font size, and other device or user setting characteristics) that can add complexity to establishing validity of results for real-world situations.

In addition, when translating to mobile use by everyday people, some aspects of evaluation may need to be simplified from studies focused on desktop or large display-based visualizations aimed at professionals. Tasks such as identifying the extreme values (e.g., max, min) and trends are perhaps not the most important things to understand about a visualization in a mobile application intended for the mass-market. Instead, tasks for mobile visualization are more personal and about users being able to understand what the visualization means for their life over understanding raw numbers, identifying trends, or statistics.

We hope that researchers will use our framework and progression of goals in order to better align their evaluation with the stage of their research. By answering more basic questions first such as perception and feasibility, researchers can fail faster and earlier in the process by understanding basic deficiencies in design before progressing to longer multi-week field studies. This framework can also help researchers to better tailor research methods to the specific stage of their development and types of research questions that they seek to answer.

Finally, there are new considerations to take into account when conducting mobile visualization research. The first is the participants that you choose to recruit. For example, if your design is meant to be used by wider populations, ensuring that your participants have a variety of educational backgrounds and that they are varied in age and are gender-balanced can help ensure that your design can be used by all. Whereas traditional visualization evaluations typically employ targeted recruiting (e.g., those with a certain level of visualization literacy, college/graduate students, data analysts), recruiting a broader spectrum of user groups may be important for evaluating mobile applications that convey quantitative information pertaining to weather, news, personal health, finance, and other topics of pertinence to the wider population.

Second, there may be additional safety and human subjects considerations to take into account. If you are collecting data over multiple weeks from a user's own handset, laws such as GDPR [35] and CCPA [5] might apply. You may need to make sure data is anonymized and encrypted in transit and stored in a safe location with limited access. The data you are collecting might also be sensitive, potentially including health or location data of the user. Participants' physical safety may be another concern if you are having them use your design while walking or exercising.

However, regardless of these potential difficulties, we would like to stress that real world conditions are so different from a lab, and that learnings from evaluation in real contexts are so important that giving up the control of a lab experiment is often the best course. Having real users engage with a visualization using their own data in their own settings on their own devices can show you if your visualization is valid for real people in the world, which is much more powerful than stating that it only works in a very controlled lab setting for a very biased subset of the population.

We hope that this framework of six evaluation goals can help researchers to pick the best methods for their goals, and to answer research questions related to goals in a logical order as their project progresses.

REFERENCES

[1] Amini, F., Hasan, K., Bunt, A., and Irani, P. "Data Representations for In-Situ Exploration of Health and Fitness Data". In: *Proceedings of the Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. ACM, 2017, pp. 163–172 (cited on page 187).

[2] Barnard, L., Yi, J. S., Jacko, J. A., and Sears, A. "Capturing the Effects of Context on Human Performance in Mobile Computing Systems". In: *Personal and Ubiquitous Computing* 11.2 (2007), pp. 81–96 (cited on page 188).

[3] Barrett, L. F. and Barrett, D. J. "An Introduction to Computerized Experience Sampling in Psychology". In: *Social Science Computer Review* 19.2 (2001), pp. 175–185 (cited on page 182).

[4] Baudisch, P. and Rosenholtz, R. "Halo: A Technique for Visualizing Off-Screen Objects". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. **Open Access version:** https://www.researchgate.net/publication/2557993_Halo_a_Technique_for_Visualizing_Off-Screen_Locations. ACM, 2003, pp. 481–488. DOI: 10.1145/642611.642695 (cited on page 186).

[5] Becerra, X. *California Consumer Privacy Act (CCPA)*. Website. Accessed: 2020-03-17. 2019. URL: https://oag.ca.gov/privacy/ccpa (cited on page 197).

[6] Bentley, F. and Tollmar, K. "The Power of Mobile Notifications to Increase Wellbeing Logging Behavior". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI '13. Paris, France: ACM, 2013, pp. 1095–1098. DOI: 10.1145/2470654.2466140. URL: https://doi.org/10.1145/2470654.2466140 (cited on page 190).

[7] Bentley, F., Tollmar, K., Stephenson, P., Levy, L., Jones, B., Robertson, S., Price, E., Catrambone, R., and Wilson, J. "Health Mashups: Presenting Statistical Patterns Between Wellbeing Data and Context in Natural Language to Promote Behavior Change". In: *Transactions on Computer-Human Interaction (TOCHI)* 20.5 (Nov. 2013). DOI: 10.1145/2503823 (cited on pages 184, 190).

[8] Bentley, F. R. and Metcalf, C. J. "Sharing Motion Information With Close Family and Friends". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI '07. San Jose, California, USA: ACM, 2007, pp. 1361–1370. DOI: 10.1145/1240624.1240831. URL: https://doi.org/10.1145/1240624.1240831 (cited on page 184).

[9] Blascheck, T., Besançon, L., Bezerianos, A., Lee, B., and Isenberg, P. "Glanceable Visualization: Studies of Data Comparison Performance on Smartwatches". In: *Transactions on Visualization and Computer Graphics (TVCG)* 25.1 (Jan. 2018). **Open Access version:** https://hal.inria.fr/hal-01851306, pp. 630–640. DOI: 10.1109/TVCG.2018.2865142 (cited on pages 184, 186, 187).

[10] Blumenstein, K., Niederer, C., Wagner, M., Schmiedl, G., Rind, A., and Aigner, W. "Evaluating Information Visualization on Mobile Devices: Gaps and Challenges in the Empirical Evaluation Design Space". In: *Proceedings of the Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV)*. **Open Access version:** http://mc.fhstp.ac.at/sites/default/files/publications/postprint-id11.pdf. ACM, 2016, pp. 125–132. DOI: 10.1145/2993901.2993906 (cited on page 186).

[11] Brehmer, M., Lee, B., Isenberg, P., and Choe, E. "A Comparative Evaluation of Animation and Small Multiples for Trend Visualization on Mobile Phones". In: *Transactions on Visualization and Computer Graphics (TVCG)* 26.1 (2020). **Open Access version:** https://hal.inria.fr/hal-02317687, pp. 364–374. DOI: 10.1109/TVCG.2019.2934397 (cited on pages 184, 187, 192).

[12] Brehmer, M., Lee, B., Isenberg, P., and Choe, E. "Visualizing Ranges Over Time on Mobile Phones: A Task-Based Crowdsourced Evaluation". In: *Transactions on Visualization and Computer Graphics (TVCG)* 25.1 (Jan. 2019). **Open Access version:** https://hal.inria.fr/hal-01857469, pp. 619–629. DOI: 10.1109/TVCG.2018.2865234 (cited on pages 184, 187, 192).

[13] Brehmer, M. and Munzner, T. "A Multi-Level Typology of Abstract Visualization Tasks". In: *Transactions on Visualization and Computer Graphics (TVCG)* 19.12 (Dec. 2013). **Open Access version:** http://www.cs.ubc.ca/labs/imager/tr/2013/MultiLevelTaskTypology/, pp. 2376–2385. DOI: 10.1109/TVCG.2013.124 (cited on page 188).

[14] Brown, B., Taylor, A. S., Izadi, S., Sellen, A., Kaye, J. ", and Eardley, R. "Locating Family Values: A Field Trial of the Whereabouts Clock". In: *Proceedings of the Conference on Ubiquitous Computing (Ubicomp)*. UbiComp '07. Innsbruck, Austria: Springer, 2007, pp. 354–371. DOI: 10.5555/1771592.1771613 (cited on page 184).

[15] Buering, T., Gerken, J., and Reiterer, H. "User Interaction With Scatterplots on Small Screens - A Comparative Evaluation of Geometric-Semantic Zoom and Fisheye Distortion". In: *Transactions on Visualization and Computer Graphics (TVCG)* 12.5 (Sept. 2006), pp. 829–836. DOI: 10.1109/TVCG.2006.187. URL: https://doi.org/10.1109/TVCG.2006.187 (cited on pages 184, 185).

[16] Buman, M. P., Epstein, D. R., Gutierrez, M., Herb, C., Hollingshead, K., Huberty, J. L., Hekler, E. B., Vega-López, S., Ohri-Vachaspati, P., Hekler, A. C., and Baldwin, C. M. "BeWell24: Development and Process Evaluation of a Smartphone "App" to Improve Sleep, Sedentary, and Active Behaviors in US Veterans With Increased Metabolic Risk". In: *Translational Behavioral Medicine* 6.3 (Nov. 2015), pp. 438–448. DOI: 10.1007/s13142-015-0359-3 (cited on page 184).

[17] Bureau, U. S. C. *Educational Attainment in the United States: 2018*. https://www.census.gov/data/tables/2018/demo/education-attainment/cps-detailed-tables.html. Accessed: 2020-03-17. 2019 (cited on page 179).

[18] Carpendale, S. "Evaluating Information Visualizations". In: *Information Visualization: Human-Centered Issues and Perspectives*. Edited by Kerren, A., Stasko, J. T., Fekete, J.-D., and North, C. Berlin, Heidelberg: Springer, 2008, pp. 19–45 (cited on pages 178, 180).

[19] Carter, S., Mankoff, J., Klemmer, S. R., and Matthews, T. "Exiting the Cleanroom: On Ecological Validity and Ubiquitous Computing". In: *Human-Computer Interaction* 23.1 (2008), pp. 47–99 (cited on page 195).

[20] Cheng, S. "The Research Framework of Eye-Tracking Based Mobile Device Usability Evaluation". In: *Proceedings of the Workshop on Pervasive Eye Tracking & Mobile Eye-Based Interaction (PETMEI)*. ACM, 2011, pp. 21–26. DOI: 10.1145/2029956.2029964 (cited on page 188).

[21] Chittaro, L. “Visualizing Information on Mobile Devices”. In: *Computer* 39.3 (Mar. 2006). **Open Access version:** http://hcilab.uniud.it/images/stories/publications/2006-03/VisualizingInformationMobile_IEEECOMPUTER.pdf, pp. 40–45. DOI: 10.1109/MC.2006.109 (cited on page 185).

[22] Choe, E. K., Lee, B., Kay, M., Pratt, W., and Kientz, J. A. “SleepTight: Low-Burden, Self-Monitoring Technology for Capturing and Reflecting on Sleep Behaviors”. In: *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp)*. UbiComp ’15. Osaka, Japan: ACM, 2015, pp. 121–132. DOI: 10.1145/2750858.2804266. URL: https://doi.org/10.1145/2750858.2804266 (cited on page 184).

[23] Choe, E. K., Lee, B., Munson, S., Pratt, W., and Kientz, J. “Persuasive Performance Feedback: The Effect of Framing on Self-Efficacy”. In: *American Medical Informatics Association Annual Symposium Proceedings*. **Open Access version:** https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900219/. American Medical Informatics Association, 2013, pp. 825–833 (cited on page 184).

[24] Choe, E. K., Lee, B., Zhu, H., Riche, N. H., and Baur, D. “Understanding Self-Reflection: How People Reflect on Personal Data Through Visual Data Exploration”. In: *Proceedings of the Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. ACM, 2017, pp. 173–182. DOI: 10.1145/3154862.3154881 (cited on pages 184, 193).

[25] Chung, C.-F., Wang, Q., Schroeder, J., Cole, A., Zia, J., Fogarty, J., and Munson, S. A. “Identifying and Planning for Individualized Change: Patient-Provider Collaboration Using Lightweight Food Diaries in Healthy Eating and Irritable Bowel Syndrome”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3.1 (2019), pp. 1–27 (cited on page 193).

[26] Consolvo, S., Bentley, F. R., Hekler, E. B., and Phatak, S. S. *Mobile User Research: A Practical Guide*. Morgan & Claypool, 2017. DOI: 10.2200/S00763ED1V01Y201703MPC012 (cited on page 182).

[27] Consolvo, S., Everitt, K., Smith, I., and Landay, J. A. “Design Requirements for Technologies That Encourage Physical Activity”. In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI ’06. Montréal, Québec, Canada: ACM, 2006, pp. 457–466. DOI: 10.1145/1124772.1124840. URL: https://doi.org/10.1145/1124772.1124840 (cited on page 189).

[28] Consolvo, S., Libby, R., Smith, I., Landay, J., McDonald, D., Toscos, T., Chen, M., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., and LeGrand, L. “Activity Sensing in the Wild: A Field Trial of Ubifit Garden”. In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2008, pp. 1797–1806. DOI: 10.1145/1357054.1357335 (cited on pages 184, 192).

[29] Correia, N., Rodrigues, A., Amorim, T., Hawkey, J., and Oliveira, S. "A Mobile System to Visualize Patterns of Everyday Life". In: *Proceedings of the Symposium on Ambient Intelligence*. Springer, 2011, pp. 241–245 (cited on pages 184, 190).

[30] Csikszentmihalyi, M. and Larson, R. "Validity and Reliability of the Experience-Sampling Method". In: *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*. Dordrecht: Springer, 1987, pp. 35–54. DOI: 10.1007/978-94-017-9088-8_3. URL: https://doi.org/10.1007/978-94-017-9088-8_3 (cited on page 182).

[31] Desai, P. M., Levine, M. E., Albers, D. J., and Mamykina, L. "Pictures Worth a Thousand Words: Reflections on Visualizing Personal Blood Glucose Forecasts for Individuals With Type 2 Diabetes". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI '18. Montreal QC, Canada: ACM, 2018. DOI: 10.1145/3173574.3174112. URL: https://doi.org/10.1145/3173574.3174112 (cited on page 184).

[32] Dix, A., Dix, A. J., Finlay, J., Abowd, G. D., and Beale, R. *Human-Computer Interaction*. Pearson Education, 2003 (cited on page 188).

[33] Eichmann, P., Edge, D., Evans, N., Lee, B., Brehmer, M., and White, C. "Orchard: Exploring Multivariate Heterogeneous Networks on Mobile Phones". In: *Computer Graphics Forum* 39.3 (2020), pp. 115–126. DOI: 10.1111/cgf.13967 (cited on page 184).

[34] Elder, J. and Zucker, S. "The Effect of Contour Closure on the Rapid Discrimination of Two-Dimensional Shapes". In: *Vision Research* 33.7 (1993), pp. 981–991 (cited on page 186).

[35] European Commission. *EU Data Protection Rules*. Law. Accessed: 2020-03-17. 2018. URL: https://ec.europa.eu/info/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules/eu-data-protection-rules_en (cited on page 197).

[36] Eyezag. *Eye Tracking on Mobile Devices*. Website. Accessed 2020-03-17. URL: https://eyezag.com/eye-tracking/mobile/ (cited on page 188).

[37] Fortmann, J., Cobus, V., Heuten, W., and Boll, S. "WaterJewel: Design and Evaluation of a Bracelet to Promote a Better Drinking Behaviour". In: *Proceedings of the Conference on Mobile and Ubiquitous Multimedia (MUM)*. MUM '14. Melbourne, Victoria, Australia: ACM, 2014, pp. 58–67. DOI: 10.1145/2677972.2677976 (cited on page 184).

[38] Froehlich, J., Dillahunt, T., Klasnja, P., Mankoff, J., Consolvo, S., Harrison, B., and Landay, J. A. "UbiGreen: Investigating a Mobile Tool for Tracking and Supporting Green Transportation Habits". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI '09. Boston, MA, USA: ACM, 2009, pp. 1043–1052. DOI: 10.1145/1518701.1518861. URL: https://doi.org/10.1145/1518701.1518861 (cited on page 184).

[39] Galesic, M. and Garcia-Retamero, R. "Graph Literacy: A Cross-Cultural Comparison". In: *Medical Decision Making* 31.3 (2011), pp. 444–457 (cited on pages 179, 191).

[40] Gouveia, R., Pereira, F., Karapanos, E., Munson, S., and Hassenzahl, M. "Exploring the Design Space of Glanceable Feedback for Physical Activity Trackers". In: *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp)*. ACM, 2016, pp. 144–155 (cited on page 195).

[41] Gouveia, R., Pereira, F., Karapanos, E., Munson, S. A., and Hassenzahl, M. "Exploring the Design Space of Glanceable Feedback for Physical Activity Trackers". In: *Proceedings of the Joint Conference on Pervasive and Ubiquitous Computing (Ubicomp)*. UbiComp '16. Heidelberg, Germany: ACM, 2016, pp. 144–155. DOI: 10.1145/2971648.2971754 (cited on page 184).

[42] Gustafson, S., Baudisch, P., Gutwin, C., and Irani, P. "Wedge: Clutter-Free Visualization of Off-Screen Locations". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2008, pp. 787–796 (cited on page 186).

[43] Gustafson, S. G. and Irani, P. P. "Comparing Visualizations for Tracking Off-Screen Moving Targets". In: *Extended Abstracts of the Conference on Human Factors in Computing System (CHI)*. ACM, 2007, pp. 2399–2404 (cited on page 186).

[44] Hekler, E. B., Klasnja, P., Froehlich, J. E., and Buman, M. P. "Mind the Theoretical Gap: Interpreting, Using, and Developing Behavioral Theory in HCI Research". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI '13. Paris, France: ACM, 2013, pp. 3307–3316. DOI: 10.1145/2470654.2466452. URL: https://doi.org/10.1145/2470654.2466452 (cited on page 182).

[45] Hincapié-Ramos, J. D. and Irani, P. "CrashAlert: Enhancing Peripheral Alertness for Eyes-Busy Mobile Interaction While Walking". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2013, pp. 3385–3388. DOI: 10.1145/2470654.2466463 (cited on page 188).

[46] Holtzblatt, K. and Beyer, H. *Contextual Design: Defining Customer-Centered Systems*. Elsevier, 1997 (cited on page 182).

[47] Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., and Möller, T. "A Systematic Review on the Practice of Evaluating Visualization". In: *Transactions on Visualization and Computer Graphics (TVCG)* 19.12 (2013). **Open Access version:** https://hal.inria.fr/hal-00846775, pp. 2818–2827. DOI: 10.1109/TVCG.2013.126 (cited on page 181).

[48] Kalnikaite, V., Rogers, Y., Bird, J., Villar, N., Bachour, K., Payne, S., Todd, P. M., Schöning, J., Krüger, A., and Kreitmayer, S. "How to Nudge in Situ: Designing Lambent Devices to Deliver Salient Information in Supermarkets". In: *Proceedings of the Conference on Ubiquitous Computing (Ubicomp)*. UbiComp

'11. Beijing, China: ACM, 2011, pp. 11–20. DOI: 10.1145/2030112.2030115 (cited on pages 184, 194).

[49] Kane, S. K., Wobbrock, J. O., and Smith, I. E. "Getting Off the Treadmill: Evaluating Walking User Interfaces for Mobile Devices in Public Spaces". In: *Proceedings of the Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*. ACM, 2008, pp. 109–118 (cited on page 188).

[50] Karkar, R., Schroeder, J., Epstein, D. A., Pina, L. R., Scofield, J., Fogarty, J., Kientz, J. A., Munson, S. A., Vilardaga, R., and Zia, J. "TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. **Open Access version:** https://depstein.net/assets/pubs/rkarkar_chi17.pdf. Denver, Colorado, USA: ACM, 2017, pp. 6850–6863. DOI: 10.1145/3025453.3025480 (cited on pages 184, 193).

[51] Kay, M., Choe, E. K., Shepherd, J., Greenstein, B., Watson, N., Consolvo, S., and Kientz, J. A. "Lullaby: A Capture & Access System for Understanding the Sleep Environment". In: *Proceedings of the Conference on Ubiquitous Computing (Ubicomp)*. UbiComp '12. **Open Access version:** https://www.researchgate.net/publication/261849983. Pittsburgh, Pennsylvania: ACM, 2012, pp. 226–234. DOI: 10.1145/2370216.2370253 (cited on page 184).

[52] Kay, M., Kola, T., Hullman, J. R., and Munson, S. A. "When (Ish) Is My Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI '16. San Jose, California, USA: ACM, 2016, pp. 5092–5103. DOI: 10.1145/2858036.2858558. URL: https://doi.org/10.1145/2858036.2858558 (cited on pages 184, 192).

[53] Kim, S. and Paulos, E. "InAir: Measuring and Visualizing Indoor Air Quality". In: *Proceedings of the Conference on Ubiquitous Computing (Ubicomp)*. UbiComp '09. Orlando, Florida, USA: ACM, 2009, pp. 81–84. DOI: 10.1145/1620545.1620557. URL: https://doi.org/10.1145/1620545.1620557 (cited on page 184).

[54] Kim, Y.-H., Jeon, J. H., Lee, B., Choe, E. K., and Seo, J. "OmniTrack: A Flexible Self-Tracking Approach Leveraging Semi-Automated Tracking". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1.3 (Sept. 2017). **Open Access version:** https://omnitrack.github.io/assets/files/IMWUT-2017-Kim-OmniTrack.pdf. DOI: 10.1145/3130930 (cited on pages 184, 190).

[55] King, A. C., Hekler, E. B., Grieco, L. A., Winter, S. J., Sheats, J. L., Buman, M. P., Banerjee, B., Robinson, T. N., and Cirimele, J. "Harnessing Different Motivational Frames via Mobile Phones to Promote Daily Physical Activity and Reduce Sedentary Behavior in Aging Adults". In: *PLOS ONE* 8.4 (2013), e62613 (cited on page 194).

[56] Klasnja, P., Consolvo, S., and Pratt, W. “How to Evaluate Technologies for Health Behavior Change in HCI Research”. In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI ’11. Vancouver, BC, Canada: ACM, 2011, pp. 3063–3072. DOI: 10.1145/1978942.1979396 (cited on page 196).

[57] Klasnja, P., Smith, S., Seewald, N. J., Lee, A., Hall, K., Luers, B., Hekler, E. B., and Murphy, S. A. “Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-Randomized Optimization Trial of HeartSteps”. In: *Annals of Behavioral Medicine* 53.6 (2019), pp. 573–582 (cited on page 195).

[58] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S. “Empirical Studies in Information Visualization: Seven Scenarios”. In: *Transactions on Visualization and Computer Graphics (TVCG)* 18.9 (Sept. 2012). **Open Access version:** https://hal.inria.fr/hal-00932606, pp. 1520–1536. DOI: 10.1109/TVCG.2011.279 (cited on page 181).

[59] Langner, R., Horak, T., and Dachselt, R. “VisTiles: Coordinating and Combining Co-Located Mobile Devices for Visual Data Exploration”. In: *Transactions on Visualization and Computer Graphics (TVCG)* 24.1 (Jan. 2018). **Open Access version:** https://imld.de/cnt/uploads/Langner_VisTiles_InfoVis17.pdf, pp. 626–636. DOI: 10.1109/TVCG.2017.2744019 (cited on pages 184, 188, 189).

[60] Lazar, J., Feng, J., and Hochheiser, H. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann, 2017 (cited on page 179).

[61] Lin, J. J., Mamykina, L., Lindtner, S., Delajoux, G., and Strub, H. B. “Fish’n’Steps: Encouraging Physical Activity With an Interactive Computer Game”. In: *Proceedings of the Conference on Ubiquitous Computing (Ubicomp)*. Edited by Dourish, P. and Friday, A. Berlin, Heidelberg: Springer, 2006, pp. 261–278. DOI: 10.1007/11853565_16 (cited on page 184).

[62] Lin, M., Goldman, R., Price, K. J., Sears, A., and Jacko, J. “How Do People Tap When Walking? An Empirical Investigation of Nomadic Data Entry”. In: *International Journal of Human-Computer Studies* 65.9 (2007), pp. 759–769 (cited on page 188).

[63] Mizobuchi, S., Chignell, M., and Newton, D. “Mobile Text Entry: Relationship Between Walking Speed and Text Input Task Difficulty”. In: *Proceedings of the Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*. ACM, 2005, pp. 122–128 (cited on page 188).

[64] Neshati, A., Sakamoto, Y., Leboe-McGowan, L. C., Leboe-McGowan, J., Serrano, M., and Irani, P. “G-Sparks: Glanceable Sparklines on Smartwatches”. In: *Proceedings of the Graphics Interface Conference (GI)*. **Open Access version:** https://doi.org/10.20380/GI2019.23. Kingston, Canada: Canadian Human-Computer Communications Society, 2019. DOI: 10.20380/GI2019.23 (cited on pages 184, 186).

[65] Nielsen, J. *Usability Metrics*. https://www.nngroup.com/articles/usability-metrics/. Accessed: 2020-03-17. 2001 (cited on page 188).

[66] North, C. "Toward Measuring Visualization Insight". In: *Computer Graphics and Applications (CG&A)* 26.3 (May 2006), pp. 6–9. DOI: 10.1109/MCG.2006.70. URL: http://dx.doi.org/10.1109/MCG.2006.70 (cited on pages 193, 194).

[67] Olson, J. S. and Kellogg, W. A. *Ways of Knowing in HCI*. Springer, 2014 (cited on page 179).

[68] Palen, L. and Salzman, M. "Voice-Mail Diary Studies for Naturalistic Data Capture Under Mobile Conditions". In: *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW)*. CSCW '02. New Orleans, Louisiana, USA: ACM, 2002, pp. 87–95. DOI: 10.1145/587078.587092. URL: https://doi.org/10.1145/587078.587092 (cited on page 182).

[69] Peck, E. M., Ayuso, S. E., and El-Etr, O. "Data Is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI '19. Glasgow, Scotland Uk: ACM, 2019. DOI: 10.1145/3290605.3300474. URL: https://doi.org/10.1145/3290605.3300474 (cited on page 191).

[70] Plaisant, C. "The Challenge of Information Visualization Evaluation". In: *Proceedings of the Conference on Advanced Visual Interfaces (AVI)*. AVI '04. Gallipoli, Italy: ACM, 2004, pp. 109–116. DOI: 10.1145/989863.989880. URL: http://doi.acm.org/10.1145/989863.989880 (cited on pages 178, 180).

[71] Ploderer, B., Reitberger, W., Oinas-Kukkonen, H., and van Gemert-Pijnen, J. "Social Interaction and Reflection for Behaviour Change". In: *Personal and Ubiquitous Computing* 18 (2014), pp. 1667–1676. DOI: 10.1007/s00779-014-0779-y (cited on page 193).

[72] Pro, T. *Tobii Pro Glasses 2*. Product. Accessed 2020-03-17. URL: https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/ (cited on page 188).

[73] Riegelsberger, J. and Nakhimovsky, Y. "Seeing the Bigger Picture: A Multi-Method Field Trial of Google Maps for Mobile". In: *Extended Abstracts of the Conference on Human Factors in Computing System (CHI)*. CHI EA '08. Florence, Italy: ACM, 2008, pp. 2221–2228. DOI: 10.1145/1358628.1358655. URL: https://doi.org/10.1145/1358628.1358655 (cited on page 184).

[74] Rodgers, J. and Bartram, L. "Exploring Ambient and Artistic Visualization for Residential Energy Use Feedback". In: *Transactions on Visualization and Computer Graphics (TVCG)* 17.12 (Dec. 2011), pp. 2489–2497. DOI: 10.1109/TVCG.2011.196 (cited on page 184).

[75] Rooksby, J., Rost, M., Morrison, A., and Chalmers, M. "Pass the Ball: Enforced Turn-Taking in Activity Tracking". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI '15. Seoul, Republic of Korea: ACM, 2015, pp. 2417–2426. DOI: 10.1145/2702123.2702577. URL: https://doi.org/10.1145/2702123.2702577 (cited on page 184).

[76] Roto, V., Popescu, A., Koivisto, A., and Vartiainen, E. "Minimap: A Web Page Visualization Method for Mobile Phones". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. CHI '06. Montréal, Québec, Canada: ACM, 2006, pp. 35–44. DOI: 10.1145/1124772.1124779. URL: https://doi.org/10.1145/1124772.1124779 (cited on page 184).

[77] Sadana, R., Agnihotri, M., and Stasko, J. *Touching Data: A Discoverability-Based Evaluation of a Visualization Interface for Tablet Computers*. ArXiv preprint 1806.06084. **Open Access version:** https://arxiv.org/abs/1806.06084. ArXiV, Aug. 2018 (cited on pages 184, 192).

[78] Sadana, R. and Stasko, J. "Designing Multiple Coordinated Visualizations for Tablets". In: *Computer Graphics Forum* 35.3 (June 2016). **Open Access version:** https://www.cc.gatech.edu/~stasko/papers/eurovis16-mcv.pdf, pp. 261–270. DOI: 10.1111/cgf.12902 (cited on page 192).

[79] Saito, T., Miyamura, H. N., Yamamoto, M., Saito, H., Hoshiya, Y., and Kaseda, T. "Two-Tone Pseudo Coloring: Compact Visualization for One-Dimensional Data". In: *Proceedings of the Symposium on Information Visualization (InfoVis)*. IEEE, 2005, pp. 173–180 (cited on page 186).

[80] Saket, B., Endert, A., and Stasko, J. "Beyond Usability and Performance: A Review of User Experience-Focused Evaluations in Visualization". In: *Proceedings of the Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV)*. BELIV '16. **Open Access version:** https://www.cc.gatech.edu/~john.stasko/papers/beliv16-ux.pdf. 2016, pp. 133–142. DOI: 10.1145/2993901.2993903 (cited on page 181).

[81] Saraiya, P., North, C., and Duca, K. "An Insight-Based Methodology for Evaluating Bioinformatics Visualizations". In: *Transactions on Visualization and Computer Graphics (TVCG)* 11.4 (2005), pp. 443–456 (cited on page 181).

[82] Schmidt, A., Holleis, P., Häkkilä, J., Rukzio, E., and Atterer, R. "Mobile Phones as Tool to Increase Communication and Location Awareness of Users". In: *Proceedings of the Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI)*. Mobility '06. Bangkok, Thailand: ACM, 2006, 21–es. DOI: 10.1145/1292331.1292355. URL: https://doi.org/10.1145/1292331.1292355 (cited on page 184).

[83] Schwab, M., Hao, S., Vitek, O., Tompkin, J., Huang, J., and Borkin, M. A. "Evaluating Pan and Zoom Timelines and Sliders". In: *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, 2019, pp. 1–12 (cited on page 193).

[84] Shneiderman, B. and Plaisant, C. "Strategies for Evaluating Information Visualization Tools: Multi-Dimensional In-Depth Long-Term Case Studies". In: *Proceedings of the Workshop on BEyond Time and Errors: Novel Evaluation Methods for Information Visualization (BELIV)*. BELIV '06. New York, NY, USA: ACM, 2006, pp. 1–7. DOI: 10.1145/1168149.1168158. URL: http://doi.acm.org/10.1145/1168149.1168158 (cited on page 181).

[85] Tollmar, K., Bentley, F., and Viedma, C. "Mobile Health Mashups: Making Sense of Multiple Streams of Wellbeing and Contextual Data for Presentation on a Mobile Device". In: *Proceedings of the Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 2012, pp. 65–72. DOI: 10.4108/icst.pervasivehealth.2012.248698 (cited on pages 184, 190).

[86] Tudor-Locke, C., Sisson, S. B., Lee, S. M., Craig, C. L., Plotnikoff, R. C., and Bauman, A. "Evaluation of Quality of Commercial Pedometers". In: *Canadian Journal of Public Health/Revue Canadienne de Sante'e Publique* 97 (2006), S10–S15 (cited on page 189).

[87] Tufte, E. R. *Beautiful Evidence*. Graphics Press, 2006 (cited on page 186).

[88] Wall, E., Agnihotri, M., Matzen, L., Divis, K., Haass, M., Endert, A., and Stasko, J. "A Heuristic Approach to Value-Driven Evaluation of Visualizations". In: *Transactions on Visualization and Computer Graphics (TVCG)* 25.1 (Jan. 2019). **Open Access version:** https://www.cc.gatech.edu/~john.stasko/papers/infovis18-icet.pdf, pp. 491–500. DOI: 10.1109/TVCG.2018.2865146 (cited on page 181).

[89] Wang, Y., Segal, A., Klatzky, R., Keefe, D. F., Isenberg, P., Hurtienne, J., Hornecker, E., Dwyer, T., and Barrass, S. "An Emotional Response to the Value of Visualization". In: *Computer Graphics and Applications (CG&A)* 39.5 (2019), pp. 8–17 (cited on page 181).

[90] Ware, C. *Information Visualization: Perception for Design*. English. 3rd. Morgan Kaufmann, 2012 (cited on page 185).

[91] Whitlock, M., Wu, K., and Szafir, D. A. "Designing for Mobile and Immersive Visual Analytics in the Field". In: *Transactions on Visualization and Computer Graphics (TVCG)* 26.1 (Jan. 2020). **Open Access version:** https://arxiv.org/abs/1908.00680, pp. 503–513. DOI: 10.1109/TVCG.2019.2934282 (cited on page 190).