Evaluating the Potential of Data-Driven Surveys for Fitness-Tracking Research

LEV VELYKOIVANENKO, University of Lausanne, Switzerland
KAVOUS SALEHZADEH NIKSIRAT, Max Planck Institute for Security and Privacy, Germany
AMRO ABDRABO, University of Lausanne, Switzerland
KÉVIN HUGUENIN, University of Lausanne, Switzerland

Online surveys are used extensively in fitness-tracking research. One common limitation is that researchers cannot assess the veracity of participants' self-reported physical activity data. One promising response to this limitation is using data-driven surveys that integrate participants' online account data (e.g., Fitbit). In this paper, we evaluate using data-driven surveys for fitness-tracking research by examining how participants perceive them, how participants' self-reported data compares to their Fitbit data, and what monetary incentives could motivate participation. To this end, we integrated the participants' Fitbit data in a survey and conducted a three-group study with $N = 300$ participants. We discuss the two main findings: First, although participants were receptive to data-driven surveys, the groups that had to share their data had lower completion rates. Second, there was a large discrepancy between the self-reported data and the data in the participants' Fitbit accounts (e.g., only 38% self-reported a consistent typical weekly exercise-time). We provide suggestions for researchers who conduct online surveys, including data-driven ones, that collect online-account data.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; **Empirical studies in ubiquitous and mobile computing**; *Empirical studies in HCI*.

Additional Key Words and Phrases: online surveys, data-driven surveys, online accounts, fitness tracking, wearable, privacy.

ACM Reference Format:
Lev Velykoivanenko, Kavous Salehzadeh Niksirat, Amro Abdrabo, and Kévin Huguenin. 2025. Evaluating the Potential of Data-Driven Surveys for Fitness-Tracking Research. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 215 (December 2025), 40 pages. https://doi.org/10.1145/3770665

1 INTRODUCTION

Fitness tracking has many reported benefits and is increasingly popular [84]. It has been studied extensively in the literature [84, 95]. For example, recent user studies have investigated usage patterns and users' utility perceptions [1, 24, 94, 96, 108], and best-practices for creating fitness-data open-datasets [62]. Large-scale fitness-data collection has been used to study how nation-level fitness campaigns affect physical activity levels [64]. The privacy risks posed by fitness trackers have also been studied [36, 109–111]. Other lines of research focus on the devices themselves, for example, by studying their security [19, 74] or their accuracy [23, 33, 39].

A common research method for studying physical activity and fitness-tracking practices is online user surveys [95]. They are often used to study user perception regarding fitness tracking (e.g., utility and privacy [94, 96, 108]) and to collect *self-reported* data about their *tracked* physical activity that can be studied or

Authors' Contact Information: Lev Velykoivanenko, University of Lausanne, Lausanne, Switzerland, lev.velykoivanenko@unil.ch; Kavous Salehzadeh Niksirat, Max Planck Institute for Security and Privacy, Bochum, Germany, kavous.salehzadehniksirat@unil.ch; Amro Abdrabo, University of Lausanne, Lausanne, Switzerland, amro.abdrabo@unil.ch; Kévin Huguenin, University of Lausanne, Lausanne, Switzerland, kevin.huguenin@unil.ch.

ACM 2474-9567/2025/12-ART215
https://doi.org/10.1145/3770665

used to screen and/or characterize participants [26]. However, online surveys also present several limitations, including biased responses [91] and concerns about data quality [37, 101] that can stem from reduced attention or engagement [67]. For example, some types of information are more salient and easier to recall, such as which physical exercise activities one engages in regularly, whereas other details, such as the number of steps taken on a certain date, are harder to recall (i.e., recall bias [10, 60]). Additionally, to present themselves more favorably, some participants can over-report healthy behaviors such as claiming they exercise more frequently than they actually do (i.e., social desirability bias [43, 91]).

One way to address such issues is through direct data collection from the participants' fitness-tracking accounts (e.g., data donation, APIs, web scrapping). This data can be used for survey *personalization* that is customizing the survey for each participant, such as modifying the survey structure [21, 22]. Survey personalization can increase completion rates and improve participant engagement [42], but its effects on data quality are not entirely clear [48, 59]. To personalize surveys at scale, custom solutions are often required [84]. For example, some researchers extend survey platforms to incorporate behavioral data [27, 52]. Recently, a system called 'Data-Driven Surveys' (DDS) [95] which enables survey personalization by integrating participant online-account data, was released. For example, it enables using data from fitness-tracking platforms to streamline screening and to collect physical-activity data.

While tools like DDS enable streamlining and personalizing surveys using data from online accounts, how participants perceive the usability, utility, and privacy of such data-driven approaches is not well understood—especially in the context of *fitness tracking*. The gap is important, as privacy concerns can reduce participants' willingness to engage with these tools, thereby influencing response rates and data quality. In particular, two important questions arise: Are participants *willing to share their data directly*, instead of manually answering survey questions? And, what are the *benefits and shortcomings* of such an approach? In this paper, we investigate the potential of data-driven surveys as a form of automatic data-collection and online survey personalization for fitness-tracking research. To this end, we use DDS as a *case study* to achieve in-depth survey personalization and to understand the extent to which participants are willing to engage with data-driven surveys, particularly given the privacy implications of sharing personal data. We pose the following research questions (RQs):

RQ1 How do participants perceive the usability, utility, and privacy of data-driven surveys in the context of fitness tracking?

RQ2 What discrepancies exist between self-reported and actual fitness data, and how do participants perceive these discrepancies?

RQ3 What value of additional monetary incentive is required to motivate participants to share their fitness data for use in data-driven surveys?

To address our RQs, we conducted an online survey by using DDS to create a data-driven survey on Qualtrics targeted at Fitbit users. The participants were randomly assigned to one of the three experimental groups: (1) the first group completed a standard, self-reported survey (i.e., henceforth ● MANUAL), (2) the second group had to share their Fitbit data and completed a shorter, personalized survey (i.e., henceforth ● DATA-DRIVEN), and (3) the third group had to both share their data and to complete the same survey as the first group (i.e., henceforth ● HYBRID). Our study uses a mixed design, including both between-subject and within-subject components, which enables us to assess both the experiential and behavioral impacts of data-driven surveys. The between-subjects design enabled us to evaluate participants' perceptions of the usability, utility, and privacy of data-driven surveys (RQ1), and to estimate the additional monetary incentive required to motivate users to share (RQ3). The within-subject design enabled us to evaluate the (in)consistency[1] between participants' self-reported fitness data with their Fitbit data (RQ2).

[1] When analyzing the (in)consistency between the self-reported data and Fitbit data we use the term 'consistent' to refer to data that is within an allowed error margin.

Our two key findings are the following. First, we found that users are generally receptive to the concept of data-driven surveys, but only about one in three are willing to share their fitness data. Furthermore, participants who had to share their data had far lower response and completion rates. Second, we found substantial discrepancies between self-reported data and Fitbit data, even when participants checked their data in their Fitbit account. For example, 27% of the participants consistently reported their most frequent activity and only 38% self-reported a consistent date for their most active day over the last six months. This means that data-driven surveys can be a better alternative to surveys relying on self-reported data. Our contributions are threefold:

- First, we provide the first empirical evaluation of participants' willingness to share their Fitbit data when participating in data-driven surveys, along with their perceptions of usability, utility, and privacy.
- Second, we analyze, in the context of a data-driven survey, the discrepancies between participants' self-reported data and the data collected from their Fitbit account.
- Third, we provide practical recommendations for researchers conducting fitness-tracking studies using data-driven surveys—covering participant recruitment, survey design, communication with participants, and budgeting.

2 RELATED WORK

2.1 Research on Survey Methodology

Online surveys, henceforth 'surveys', are an established way to conduct research in various domains [28]. Surveys offer benefits, such as reaching many participants [28, 68], cost-effectiveness [28, 68, 85], fast data collection [28, 68], collecting various types of data [11, 28, 107], and anonymity [11]. Evans and Mathur [28] compare surveys with other data collection methods, and show that surveys enable quickly reaching many participants at a low cost. Menon and Muraleedharan [68] study the relevance and methodological considerations when conducting surveys, and show that surveys offer benefits such as fast large-scale and low-cost data collection. Sammut et al. [85] study ways to increase survey response rates, and find that using personalized invitations and participation reminders can increase response rates. Braun et al. [11] explore conducting qualitative surveys, and argue that participants feeling anonymous can enable greater sharing, especially regarding sensitive topics. Zimba and Gasparyan [107] explore using surveys as a data collection method in medical and public health research, and show that surveys enable collecting relevant data quickly, especially in rapidly changing fields.

Surveys have multiple (potential) downsides such as biased responses [91], causing survey fatigue [29, 77, 99], limited generalizability [3], low response rates [100], low completion rates [65], and low data quality [37, 101]. Steenkamp et al. [91] study social desirability bias in survey responses. They find that socially desirable responses can introduce unnecessary variation in scale measures. Fass-Holmes [29] conducts a systematic literature review of survey fatigue. They find that research on survey fatigue is limited, survey fatigue is not defined consistently, and that there is an unclear effect of survey fatigue on response rates. Andrade [3] studies the methodological limitations of surveys. They find that participant populations are difficult to describe, which can lead to limited generalizability, and that self-selection bias among participants can distort survey results. Wu et al. [100] conduct a meta-analysis of education-related studies using surveys. They find that surveys had an average response rate of 44.1%, which was primarily caused by researchers not targeting the most relevant populations for their research. Liu and Wronski [65] study survey completion rates. They find that survey length is inversely related to completion rates, with higher proportions of open-ended questions lowering completion rates the most. Gadiraju et al. [37] studied the quality of survey responses obtained from online crowdsourcing platforms. They find that, depending on the country, a significant proportion of participants provide low-quality answers. All the aforementioned studies provide guidelines to handle the issues that they explored. Nonetheless, innovations in survey design and research tools could also address these issues and improve survey quality.

2.2 Innovations in Survey Tools

There are many survey platforms, including Qualtrics, Google Forms, SurveyMonkey, and LimeSurvey, to name the most popular ones among researchers [95]. Researchers explore extending the most popular survey platforms with new features, which enable creating novel questions and experiments [14, 25, 69, 95, 99], and integrating participants' online account data [95]. Wen and Colley [99] propose creating a system that combines interviews and standard surveys to monitor participant responses and to open a chat with a given participant to ask them to provide more details to their answers in open-ended questions. Celino and Re Calegari [16] developed CONEY, a toolkit for creating conversational surveys, showing that participants preferred the interactive surveys over conventional ones. Molnar [69] developed SMARTRIQS, a toolkit for creating interactive online experiments on Qualtrics. Ebert et al. [25] created QButterfly, a toolkit for conducting user interaction studies on Qualtrics and LimeSurvey. Carpenter et al. [14] created templated custom HTML and JavaScript questions to conduct implicit association tests on Qualtrics.

Recently, Velykoivanenko et al. [95] created DDS to integrate participants' online account data in Qualtrics and SurveyMonkey surveys. DDS can extract specific data and import it as embedded data in surveys; the extracted data remains available to the researchers, for example, for studying or characterizing the participants. We extended and used DDS for our experiment (see Section 3.2 for more details).

There are also several platforms and frameworks for collecting participants' data from smartphones and wearable devices. AWARE[2] is a framework for Android [34, 92] and iOS [72, 73] smartphone apps that enables experience sampling and collecting smartphone sensor data. Digital Biomarker Discovery Pipeline (DBDP) [7] is a framework for discovering digital biomarkers that are collections of data that are indicative of relevant health outcomes.[3] The DBDP project website[4] showcases various data-processing pipelines for identifying digital biomarkers. Mobile Data To Knowledge (MD2K)[5] is a research initiative centered on using mobile sensor data in health research. Modular Open Research Platform (MORE)[6] is a web and smartphone app for conducting longitudinal and intervention studies. Remote Assessment of Disease and Relapses (RADAR-base) [78][7] is a platform and framework for Android and iOS for real-time data collection from smartphone sensors, smartphone apps, and connected wearable devices. Way to Health (W2H)[8] is a web-based platform for creating behavior change interventions.

Researchers have also explored integrating chatbots in surveys to see if they could lead to more insightful responses from participants, especially in open-ended questions [55, 80, 97, 101, 105]. Kim et al. [55] compare a standard Web survey with a survey delivered through the Facebook Messenger client, in a casual and formal conversational style. They find that the more interactive and personalized Facebook Messenger survey improved the quality of participants' responses. Xiao et al. [101] compare a standard Web survey with a chatbot-powered one for a survey consisting mainly of open-ended questions. They find that the more interactive and self-personalizing chatbot-powered survey led to higher quality and more insightful responses. Zarouali et al. [105] compare using chatbots with standard Web surveys for longitudinal research, they find surveys using chatbots perform better on attention checks but worse in internal consistency. Rhim et al. [80] compare a humanized and a non-humanized chatbot. They find that the participants were more positive towards the humanized chatbot and disclosed more information to it, but their responses also showed more social desirability bias. Wei et al. [97] explore using LLM-powered chatbots to replace traditional surveys. They find that the LLM-powered chatbot covered most

[2] See: https://awareframework.com/, last visited: Oct. 2025.

[3] For example, inferring if someone has COVID-19 from fitness-tracker data.

[4] See: https://dbdp.org, last visited: Oct. 2025.

[5] See: https://md2k.org/, last visited: Oct. 2025.

[6] See: https://more-platform.at/, last visited: Oct. 2025.

[7] See: https://radar-base.org, last visited: Oct. 2025.

[8] See: https://chti.upenn.edu/way-to-health, last visited: Oct. 2025.

of the questions that they wanted to ask and had some advantages such as handling diverse responses and asking follow-up questions. However, it also had drawbacks such as producing random answers and being overly repetitive. Jacobsen et al. [53] explore using different interview probes with LLM-powered chatbots instead of traditional open-ended questions. They find that certain probes were better for collecting high-quality data at specific stages of the survey.

2.3 The Use of Surveys in Fitness-Tracking Research

Surveys have been used extensively for studying wearable activity trackers [84, 89]. In their interview-based study on IoT adoption, Page et al. [75] use a survey to collect additional information about IoT usage, including fitness-tracker usage. Velykoivanenko et al. [94] use a survey to study Fitbit users' perceptions of the privacy and utility of their Fitbit fitness-trackers. Lee et al. [62] use surveys, throughout a longitudinal study on creating open-datasets of fitness-tracker users, to collect various types of information. Li et al. [63] use a survey to collect supplemental information for their interview-based study of self-tracking practices. Epstein et al. [27] use a survey to study the utility perceptions and data presentation preferences of Fitbit users who stopped using their fitness trackers. Epstein et al. [26] use a survey to identify the reasons people stop using fitness trackers and the lasting effects that their fitness tracker usage had. Gabriele and Chiasson [36] use a survey to study fitness-tracker users' perceptions of the security and privacy aspects of fitness trackers. Zufferey et al. [110] use a survey to study fitness-tracker users' perceptions of the security and privacy aspects of sharing their fitness data with third-party applications.

2.4 Research Gap

Extensive research exists on online-survey methodology and on developing new tools for improving research quality. However, while papers presenting new tools often evaluate tool quality, they tend not to evaluate participants' privacy and utility perceptions of using such tools. Specifically, to the best of our knowledge, evaluating participant perception of data-driven survey tools that collect data from additional online services has been understudied.

3 METHODOLOGY

Our research aims to evaluate the potential of data-driven surveys as an instrument for fitness-tracking research. Our research questions (RQs) are related to two primary areas, specifically (1) participants' perceptions of data-driven surveys (RQ1) and (2) the discrepancy between participants' (physical activity-related) self-reported data and the data extracted from their Fitbit accounts (RQ2). Additionally, we assess participants' expected monetary incentive for sharing their fitness data in data-driven surveys (RQ3). To answer our RQs, we designed a user study based on an online (data-driven) survey. Figure 1 illustrates our experimental design. The study was approved by our institutional review board (IRB); for more information about the ethical considerations, see Section 3.7.

3.1 Overview and Rationale

To answer RQ1 (i.e., participant perception of data-driven surveys), we designed a 'mock' survey about fitness tracking and complemented it with questions about participants' perceptions regarding the 'mock' survey. To compare perceptions of taking a data-driven survey with those of taking a normal survey, we needed a baseline. Hence, we conducted a randomized two-arm experiment (between-subject design) with two participant groups: ● DATA-DRIVEN (experimental group) and ● MANUAL (control group). The 'mock' survey comprised two parts. In the first part of the 'mock' survey, we collected factual data related to physical activity (recorded by the participants' Fitbit device), namely categorical and numerical, such as the most frequent activity type (M.Q1) and the average weekly exercise-time (M.Q2). Only the ● MANUAL group answered the first part (factual) of the 'mock'

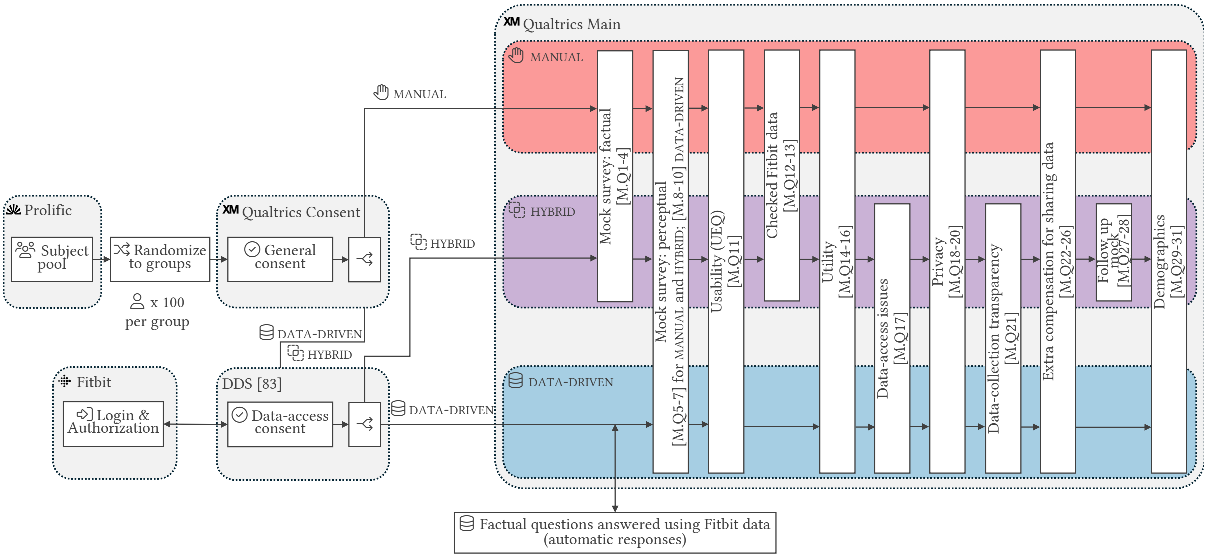

Prolific
Subject pool
Randomize to groups
x 100 per group
Qualtrics Consent
General consent
MANUAL
HYBRID
DATA-DRIVEN
HYBRID
Fitbit
Login & Authorization
DDS [83]
Data-access consent
DATA-DRIVEN
Qualtrics Main
MANUAL
HYBRID
DATA-DRIVEN
Mock survey: factual [M.Q1-4]
Mock survey: perceptual [M.Q5-7] for MANUAL and HYBRID; [M.8-10] DATA-DRIVEN
Usability (UEQ) [M.Q11]
Checked Fitbit data [M.Q12-13]
Utility [M.Q14-16]
Data-access issues [M.Q17]
Privacy [M.Q18-20]
Data-collection transparency [M.Q21]
Extra compensation for sharing data [M.Q22-26]
Follow up mock [M.Q27-28]
Demographics [M.Q29-31]
Factual questions answered using Fitbit data (automatic responses)

Fig. 1. Experimental Design Procedure.

survey. In contrast, the • DATA-DRIVEN group did not answer these questions as they were answered automatically using the data from their fitness-tracking account to which they granted us access. This process is described in Section 3.2. In the second part of the 'mock' survey, the participants elaborated on their answers to the first part of the 'mock' survey. Both groups answered these questions. To assess the participants' utility, usability, and privacy perceptions of taking a data-driven survey, we asked the same questions to both groups. However, the questions for the • MANUAL group were phrased conditionally, and the participants received an explanation with illustrations of what a data-driven survey is (e.g., M.Part 4).

To answer RQ2 (i.e., discrepancy between self-reported and Fitbit data), we needed access to *both* the self-reported *and* fitness-tracking data (within-subject design). Hence, we created a *third* participant group, called • HYBRID. The • HYBRID group was asked to grant us access to their fitness-tracking data, but also to complete the first part of the 'mock' survey. This enabled us to investigate, through follow-up questions (M.Q27-M.Q28), the reasons behind the discrepancies between their self-reported and fitness-tracking data. The • HYBRID group also completed the 'perceptions' part of the survey, which provided additional insights.

We also investigated the implications for researchers interested in conducting data-driven surveys, by analyzing the participants' behavioral data (e.g., dropout rates in all groups) and their self-reported data (e.g., additional monetary compensation for taking part in data-driven surveys; RQ3).

To conduct our study, we chose Fitbit as the fitness-tracking service provider. To implement our data-driven survey (for the • HYBRID and • DATA-DRIVEN groups), we used DDS [95]. We explain these decisions in more detail below.

3.2 Background and Instrument Selection

We chose to use Fitbit because Fitbit is one of the the most studied wearable brands [33, 39, 95]. To compare real perceptions of taking a data-driven survey with a standard survey, we needed a way to collect participants' Fitbit data. We identified three approaches for collecting actual Fitbit data, namely data donation [15, 46, 58, 98], data collection through APIs (where participants' online account data is collected before or during the survey) [2, 4, 52, 54, 95], and web scraping. Web scraping, beyond ethical issues, would not work for accessing data that is not publicly available. Hence, we had to consider data donation and collection through APIs.

Data donation consists of asking participants to request a copy of their data from online service providers[9] required for a study and to donate it to research by transferring it to the researchers conducting the study [15, 46]. Such an approach offers a way to access a significant amount of data [58, 98]. It offers participants more fine-grained control over the data they share, as they can modify the data files before sending them to researchers. However, making such modifications might be complicated for non-tech-savvy participants. Participants, hence, may unintentionally share more (or less) data than is required for a study [15]. Data donation is prone to delays and technical issues, as requesting one's data can be complicated and can take several days to be made available [15]. It was used successfully, for example, to personalize a survey on Twitter advertising [98].

An alternative to data donation is collecting data through APIs, streamlining the survey-taking process for participants by allowing them to grant access to their data instead of requesting a copy of their data. Specific parts of participants' data can then be retrieved and used to personalize their survey. This approach enables more fine-grained data collection than data donation. It was successfully used successfully, for example, to personalize surveys using Foursquare data [52] and with Facebook data [2, 4, 54].

Based on the aforementioned approaches, collecting data through APIs seemed to be the most straightforward approach and we identified DDS [95] as the most relevant tool for our study as it offers a streamlined way to integrate participants' online account data and is freely available on GitHub.[10] DDS enables integrations with online services such as Fitbit and GitHub. We extended DDS to collect some additional data that were required for our survey.

3.3 Participant Recruitment

We used Prolific to recruit participants, as its participant pool is seen as reliable [76] and it enables us to pre-screen participants for those who use wearables. We chose to recruit U.S. residents, considering the high prevalence of wearable devices among U.S. citizens [70] and that the U.S. participant pool is the largest on Prolific. We also added a Prolific pre-screening criterion to include only participants who own a fitness tracker or smartwatch. We recruited participants who (1) use Fitbit wearable devices, (2) regularly synchronize their wearables, and (3) have had a Fitbit account for at least six months. We chose participants who regularly synchronize their wearables, to make sure that they would have sufficient data in their accounts. We needed participants to have a Fitbit account for at least six months, to make sure that our participants would be comfortable with the device and app.

We began by deploying a screener survey (1 min. expected duration, GBP 0.15 (≈ USD 0.20) payment) to select participants. The screener asked for participation consent and a multiple-choice question about which fitness tracker(s) the participants use. Finally, if they selected that they use a Fitbit wearable, they were asked two follow-up questions about whether they regularly synchronize their wearable and for how long they have had a Fitbit account.

3.4 Procedure

After the screener, the selected participants were *randomly* assigned to one of three groups: (1) the ● MANUAL group, (2) the ● HYBRID group, and (3) the ● DATA-DRIVEN group. We calculated a priori the required minimal sample size by using G*Power [12, 30–32]. To conduct Kruskal-Wallis tests, Mann-Whitney tests, and proportion Z-tests (see Section 3.6 for more information) we estimated the minimum sample size for each test. For all tests, we identified a medium effect size, which resulted in setting the effect size parameter to 0.5 [12, 32]. To minimize

[9] For an example see the Google Takeout website for exporting Google data (a Google Account is required): https://takeout.google.com, last visited: Oct. 2025.

[10] See: https://github.com/DataDrivenSurveys/DataDrivenSurveys, last visited: Oct. 2025.

Type I and Type II errors, we used the standard $\alpha = 0.05$ and $1 - \beta = 0.95$. The power analyses resulted in 92 participants for Mann-Whitney tests and 87 participants for Z-tests participants per group. For the Kruskal-Wallis tests, we estimated the overall required sample size by using the parametric Analysis of Variance (ANOVA) test for three groups of participants, which resulted in a minimum of 252 participants. To allow a safety margin, we collected 100 complete and valid participant answers per group.

Figure 1 shows the overall participant flow throughout our study. The selected participants were invited to our study. Each group received a separate study invitation. All participants first had to complete a consent form (i.e., general consent) that was implemented as a separate survey (C.Part 1). The consent form had one section that changed depending on which group a participant was in. After consenting to participate, ● MANUAL participants were told that they would begin the survey and were then redirected to the main survey. After consenting to participate, ● HYBRID and ● DATA-DRIVEN participants were shown instructions about what they had to do on the DDS platform (screenshots and step-by-step instructions), then they were redirected to the DDS platform. On DDS they had to log into their Fitbit account and to grant us access to the data required for the survey as shown by the 'DDS' and 'Fitbit' blocks in Figure 1. After this, DDS downloaded the data required for the survey, processed it, uploaded it to Qualtrics, deleted it from its memory,[11] created a unique survey distribution URL, and redirected the participants to the main survey on Qualtrics that uses this URL. The ● HYBRID and ● DATA-DRIVEN groups then completed the survey as normal. The participants in all groups were paid GBP 1.50 (≈ USD 2) for participating in the study.

3.5 Survey Design

Here, we present an overview of our survey design. A detailed summary of the survey design is available in Appendix A. Also, the full transcript of the main survey is available in Appendix B. The survey consisted of two halves. First, a 'mock' survey that first had a **factual** part (M.Part 1) and then a **perceptual** part (M.Part 2). The **factual** part consisted of four questions (M.Q1-M.Q4) that collected factual information, such as how much time they spend exercising each week while wearing their Fitbit device (M.Q2). It was answered only by the ● MANUAL and ● HYBRID groups. The ● DATA-DRIVEN group's questions were "answered" using their Fitbit data. To create these questions, we referred to the types of information collected in the literature [5, 18, 35, 49, 56], that could be obtained from the Fitbit app UI, and that could be calculated using data available through the Fitbit API. The **perceptual** part consisted of three questions (M.Q5-M.Q7) that collected perceptual information related to each of the factual questions, such as what motivates the participants to exercise as much as they do (M.Q6). These questions were meant to expand on the participants' answers to the factual questions.

3.6 Data Analysis

In our results, we provide descriptive statistics for many close-ended questions (e.g., M.Q17). For some close-ended questions (e.g., M.Q14), we performed statistical analyses by following suggested best practices in HCI [103]. Hence, we used Kruskal-Wallis tests to check for significant differences between our participant groups [93]. When significant differences were found, we did Mann-Whitney tests with a Bonferroni correction as follow-up tests to identify which pairs had significant differences. To analyze the differences between the ● HYBRID groups' self-reported answers and their data we used proportion Z-tests. We further calculated Pearson's correlation coefficient to check for a link between the number of days since the most active day and the (in)consistency of participant answers. To compute the UEQ scores, we used the spreadsheet provided by the UEQ authors [50]. We followed the suggested exclusions of certain participants, based on the UEQ spreadsheet. We checked the open-answer questions for the participants who would be excluded and found them to be thoughtful, hence we excluded only the participants as identified by UEQ exclusion criteria from the UEQ analysis, but we analyzed

[11] The data is never stored on the DDS platform.

their responses in the rest of the survey. To estimate response rates, we divide the number of participants who completed the survey by the participant pool allocated to each group [13, 41]. To calculate the survey completion rates, we divide the number of participants who completed the survey by the number who started the survey [13, 41]. We analyzed the open-ended questions (e.g., M.Q27) by using inductive open coding [82]. One of the co-authors used MAXQDA to iteratively code the open-ended questions. During the process, they categorized and refined the codes. Finally, the overarching themes were identified by grouping the codes. We provide the resulting codebook in Appendix E.

3.6.1 Discrepancy Analysis. Fitbit devices have varying accuracies. Regarding activity recognition, accuracy varies significantly based on the activity being recognized,[12] ranging from 9.5% to 17.7% for biking and up to 100% for outdoor running [23]. As some participants synchronize their Fitbit accounts with third-party apps that report activities, we mapped such data to standard Fitbit activity names (e.g., we mapped 'Elliptical (MyFitnessPal)' to 'Elliptical'). We, likewise, applied the same process for participants who chose 'Other' in M.Q1 to map what they wrote to Fitbit activity names. Activities that did not have a Fitbit counterpart were recorded as what the participants wrote. Although the accuracy of Fitbit activity detection varies, since it is a categorical variable, there is no direct margin that can be applied. However, our analysis method inherently applies a margin inspired by machine learning approaches that consider certain activities as equivalent. Regarding exercise time, Germini et al. [39] find that the mean active-time error ranged from 7% to 72%, whereas Feehan et al. [33] find that it ranged from 44% to 632%. Hence, for the discrepancy analysis, we consider participants' responses to be consistent if their reported weekly exercise-time is within 50% of the value reported by Fitbit. Several recent systematic literature reviews find that Fitbit devices are reasonably accurate for measuring steps, within ±10% [33, 39]. Hence, for the analysis of the discrepancy between the participants' self-reported data and the data in their Fitbit accounts, we consider participants' responses to be consistent if their reported step-counts are within 10% of the number reported by Fitbit. For the date of the most active day, we chose to classify based on whether a participant reported a consistent date. We then further decomposed the results in order to focus on the participants who reported a date within the last 30 days before they took the survey.

It should be noted that we did not analyze the mock-survey data, beyond characterizing our sample of participants (see Table 5) and studying the discrepancy between the self-reported data and the data extracted from Fitbit for the ● HYBRID group, as it was not relevant to our work. As the data could be valuable to other researchers, to support open-science [83], we made a de-identified version of the mock-survey data available on Open Science Framework (OSF)[13], under OpenData conditions.

3.7 Ethical and Privacy Considerations

Our study was approved by our institutional review board (IRB). The participants signed a consent form that explained in detail the study participation conditions, the data that would be collected, the means that would be used for the data to be processed and stored, the way to withdraw from the study, and information about the monetary compensation for participation in the study. Participants were paid at the hourly rate suggested by Prolific. The participants who had to share their data (i.e., the ● HYBRID and ● DATA-DRIVEN groups) were given additional information in the consent form; it described the way the data that they granted us access to would be processed. In particular, they were shown, before granting access to their Fitbit data, the type of data that would be accessed and the statistics that would be computed from them (see Appendix D). The raw data accessed from Fitbit were not kept; only the statistics were kept (on Qualtrics, DDS does not retain participants' data). To prevent multiple submissions from the same Fitbit user, the Fitbit identifiers of the participants were stored only

[12] The activities that are recognized automatically are: Walk, Run, Outdoor Bike, Elliptical, Sports, Aerobic Workout, and Swimming. See Which activities does my Fitbit device track automatically?, last visited: Oct. 2025.

[13] See: http://doi.org/10.17605/osf.io/k25tn, last visited: Oct. 2025.

Table 1. Survey Participation Rates by Group

Group	Invited*	Started	Dropped-Out	Response Rate**	Completion Rate***
• MANUAL	133	105	5	75.2%	95.2%
• HYBRID	419	280	179	24.1%	36.1%
• DATA-DRIVEN	350	268	166	29.1%	38.1%

* As we do not know the exact number of participants invited by Prolific, we use the total pool of participants who were allocated to that group.

** The proportion of participants who completed the survey relative to those who were invited ($100 \times n_{\text{completed}}/n_{\text{invited}}$) [13, 41].

*** The proportion of participants who completed the survey relative to those who started it ($100 \times n_{\text{completed}}/n_{\text{started}}$) [13, 41].

on DDS for the duration of the experiment. These identifiers were ultimately deleted from DDS; they were never transferred to Qualtrics. Our methodology is closely aligned with the approaches used in data donation studies.

4 RESULTS

4.1 Participant Demographics, Survey Completion Rates, and Ratio of Checking Data

We screened a total of 3050 participants. This resulted in 902 participants meeting our screening criteria. As explained in Section 3.4, we planned to have 100 valid responses per group. Hence, we let Prolific invite participants until we reached $N = 100$ valid responses per group (i.e., they consented, completed, provided thoughtful answers to the open-ended questions; and for those in the • HYBRID and • DATA-DRIVEN groups, their Fitbit data matched the screening criteria).

Table 5 summarizes the participant demographics. The participants were diverse in terms of age, both overall and within each group. Regarding age, the overall average age was 43.5. The Kruskal-Wallis test found no significant differences between the participant groups ($\chi^2(2) = 0.4, p = .8$). Regarding gender, the • HYBRID and • DATA-DRIVEN groups were mostly women, whereas the manual group had a slight majority of men. Regarding the IUIPC score, the participants were biased towards the higher-end, as was expected [44], both overall and across the three participant groups. The Kruskal-Wallis test found no significant differences between the participant groups ($\chi^2(2) = 0.2, p = .9$). Regarding ethnicity, the participants were roughly representative of the U.S. population. The Kruskal-Wallis test found no significant differences between the groups ($\chi^2(2) = 0.23, p = .9$). To further characterize our sample, we also report some of the relevant results from the 'mock' survey. For the • HYBRID group we report both the participants' self-reported answers and the values calculated using their Fitbit data.

Table 1 summarizes the survey completion rates and estimated response rates. The completion rates for the • HYBRID and • DATA-DRIVEN groups are roughly half of those compared to the • MANUAL groups. There could be several possible explanations for this, such as privacy concerns, the complexity of the access granting process, or simply not having an actual Fitbit account.

Regarding the ratio of checking Fitbit data while answering the mock survey, a clear majority (72%) of the • MANUAL participants reported checking their Fitbit data while answering the mock survey questions (M.Q12). In contrast, nearly half (42%) of the • HYBRID participants did so (M.Q13).

4.2 Discrepancy Analysis: Self-Reported Data vs. Actual Fitbit Data

Here, we present the results related to the accuracy of the participants' self-reported answers for the • HYBRID group. For the analysis, we take the data extracted from Fitbit as the reference. However, Fitbit data can be subject to inaccuracies with respect to 'reality'. We apply the following margins when categorizing whether the participants reported values consistent with their Fitbit data: For the most frequent activity (see Section 4.2.1;

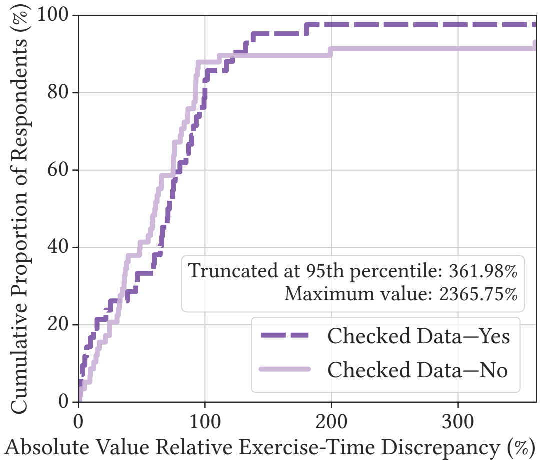


Fig. 2. Absolute Value of the Relative Exercise-Time Discrepancy ($100 \times |v_{\text{Fitbit}} - v_{\text{reported}}|/v_{\text{Fitbit}}$). “Checked Data—Yes” refers to participants who checked their Fitbit data (i.e., in the Fitbit app) before answering, and “Checked Data–No” refers to those who did not. (source: M.Q2).

M.Q1), we check if what the participants reported strictly matches what we calculated from their Fitbit data. For the exercise-time discrepancy (see Section 4.2.2; M.Q2), we consider the self-reported time consistent if it is within 50% of the value reported by Fitbit. For the most active day—date (see Section 4.2.3; M.Q3), we check if the value reported matches exactly the value from their Fitbit data. For the accuracy of step counts (see Section 4.2.4; M.Q4), we consider the self-reported time consistent if it is within 10% of the value reported by Fitbit. Section 3.6.1 provides details about how we chose these margins. The limitations of our chosen questions are discussed in Section 5.2.2.

4.2.1 Most Frequent Activity. Regarding the most frequent activity other than the ‘Walking’ activity (M.Q1), only 27% of the participants reported it consistently. Of the participants who checked their data, 31.0% reported a consistent activity, and 24.1% of those who did not check their data reported a consistent activity.

4.2.2 Weekly Exercise-Time. Figure 2 shows the CDF of the absolute value relative discrepancy between the participants' estimated weekly exercise-time and the data in their Fitbit account (M.Q2). Overall, slightly less than half (38%) of the participants reported a consistent exercise-time (i.e., an exercise time within 50% of the value reported by Fitbit). Of the participants who checked their data, 33.3% reported a consistent time and, surprisingly, 41.4% of those who did not check their data reported a consistent time. The median of the absolute value of the discrepancy is 66.2 minutes (65.6% for the absolute value of the relative discrepancy). The fact that more participants reported a consistent time, without checking their data, could be due to the Fitbit app UI making it difficult to understand the definition of average weekly exercise-time.

4.2.3 Most Active Day—Date. Overall, 26% of the participants reported a consistent date for their most active day within the last six months (M.Q3). Of the participants who checked their data, 52.4% reported a consistent date, and only 6.9% of those who did not check their data reported a consistent date. There is a significant overall correlation ($r(98) = -0.22, p = .03$) between the number of days since the most active day and whether a participant reported a consistent most active day—date. While, there is no correlation for those who checked their data, there is a significant correlation for those who did not check their data ($r(56) = -0.34, p = .01$). This means that the further back in the past the most active day was correlates with a lower chance of reporting a consistent date for participants who did not check their data.

The median of the absolute value of the discrepancy is 29 days. There is a significant positive correlation ($r(98) = 0.54, p < .001$) between the number of days since the most active day and the absolute value of the discrepancy. While there is no significant correlation for those who checked their data, there is a significant correlation for those who did not check their data ($r(56) = 0.69, p < .001$). When looking at the non-absolute value discrepancies, there are significant correlations overall ($r(98) = -0.59, p < .001$), for those who checked their data ($r(40) = -0.41, p = .007$), and for those who did not check their data ($r(56) = -0.66, p < .001$). The participants tended to report dates that were not far back enough and were closer to the date of the survey, with 50% of the participants reporting a date within 30 days of when they took the survey. Of the participants who reported a date within the last 30 days, among those who checked their data, 35.7% reported a consistent date, and 8.3% of those who did not check their data reported a consistent date. Among the participants who reported a date within the last 30 days, there is a significant correlation ($r(48) = -0.50, p < .001$) between the number of days since the most active day and reporting a date consistent with the participants' Fitbit data. While, there is no significant correlation for those who checked their data, there is a significant correlation for those who did not check their data ($r(34) = -0.42, p = .01$).

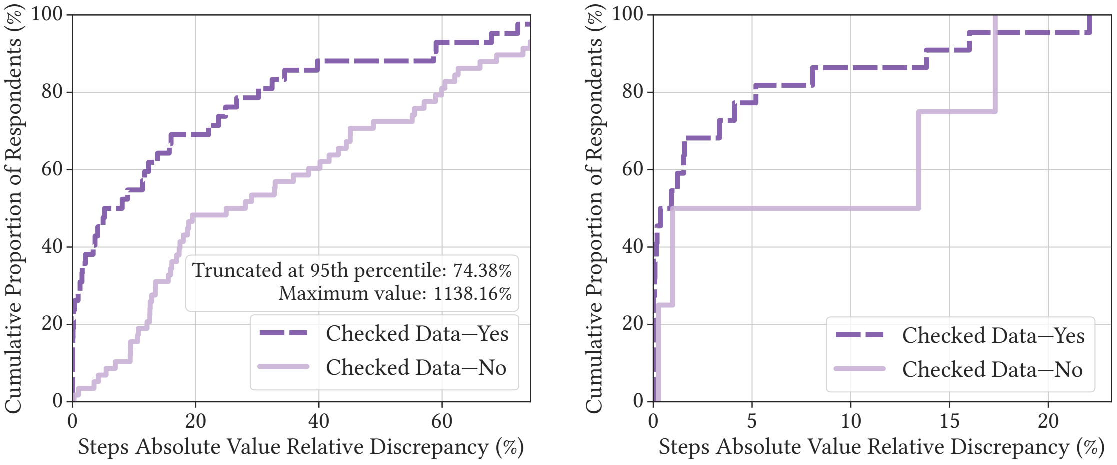
4.2.4 Most Active Day—Step Count. Figure 3a shows the CDF of the absolute value relative discrepancy between the reported and consistent number of steps (i.e., within 10% of the value reported by Fitbit) on the most active day within the last six months (M.Q4). Overall, 32% of participants reported a consistent step count. Of the participants who checked their data, 54.8% reported a consistent step count, and 15.5% of those who did not check their data reported a consistent step count. There is a significant correlation ($r(98) = -0.23, p = .02$) between the number of days since the most active date and reporting a step count consistent with the data from Fitbit. However, there were no significant correlations when looking at the participants who checked their data, nor for those who did not check their data.

The median of the absolute value of the discrepancy is 2697 steps (16.5 % for the absolute value of the relative discrepancy). There are no significant correlations between the absolute values of the step count discrepancies for the ● HYBRID group as a whole, those who checked their data, nor those who did not check their data. However, there is a weak correlation between the non-absolute values of the step count discrepancies for the ● HYBRID group as a whole ($r(98) = 0.21, p = .04$), but there are no significant correlations for those who checked their data nor those who did not check their data. This means that the further back in the past the most active day was correlates with participants reporting a lower step count than was recorded by their Fitbit.

Figure 3b shows the CDF of the absolute value of the relative discrepancy between the reported and extracted (i.e., from Fitbit) number of steps on the most active day within the last six months, but only for the participants who reported a consistent most active day—date. Overall, 80.8% of these participants reported a consistent step count. Of the participants who checked their data, 86.4% reported a consistent step count, and 50% of those who did not check their data reported a consistent step count. The median of the absolute value of the discrepancy is 147 steps.

4.2.5 Participant Perceptions of Differences between Self-Reported Answers and Fitbit Data. We asked the participants whose self-reported most frequent activity did not match their actual most frequent activity based on Fitbit data ($n = 74$, Hybrid group), to explain the discrepancy (M.Q27). The majority of these participants ($n = 48$) attributed the difference to Fitbit's limitations in accurately detecting activity types. They highlighted specific examples of misclassification: *"I would guess it is a misread, as I have not used an elliptical in the past 6 months."* [W, 23 y.o., IUIPC: 6.2];[14] *"Sometimes Fitbit counts my walking as swimming."* [W, 28 y.o., IUIPC: 4.9].

[14] This shows the participant's gender, age, and IUIPC score.



(a) Most Active Day—Steps Absolute Value of the Relative Discrepancy ($n_{\text{checked\_yes}} = 42$, $n_{\text{checked\_no}} = 58$). ($100 \times |v_{\text{Fitbit}} - v_{\text{reported}}|/v_{\text{Fitbit}}$)

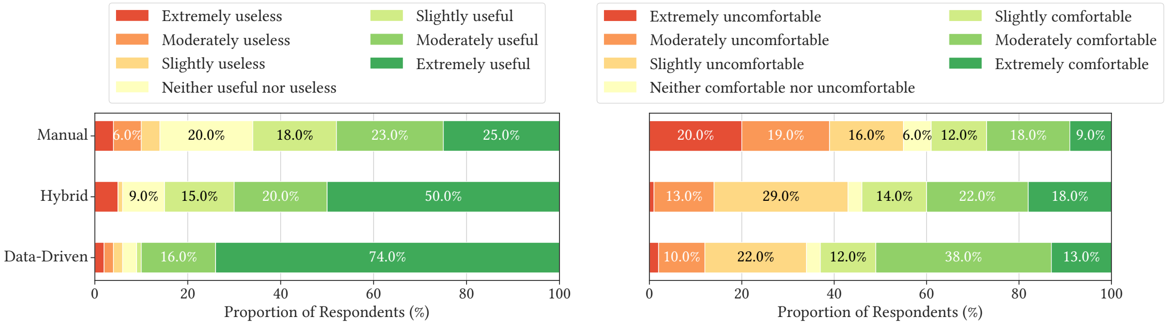
(b) Most Active Day—Steps Absolute Value of the Relative Discrepancy. Only for participants who reported a consistent date ($n_{\text{checked\_yes}} = 22$, $n_{\text{checked\_no}} = 4$). ($100 \times |v_{\text{Fitbit}} - v_{\text{reported}}|/v_{\text{Fitbit}}$)

Fig. 3. Most Active Day—Step Count Reporting Discrepancies (source: M.Q3 and M.Q4).

A smaller group of participants ($n = 7$) explained the discrepancy as a result of differing perceptions or terminology compared to Fitbit's categorization. *"I'm not very fluent in fitness vocabulary. I don't always know what to label my workouts. Sometimes, if I know I'll be working out, I'll just label them 'weights' because it seems to be the most accurate label."* [W, 25 y.o., IUIPC: 6.5]; *"Walking on the treadmill and walk is the same thing."* [W, 20 y.o., IUIPC: 6.4]. Additionally, six participants admitted to personal oversight (e.g., *"I didn't realize."* [W, 46 y.o., IUIPC: 6.9]); five mentioned forgetfulness (e.g., *"I have a horrible memory."* [W, 33 y.o., IUIPC: 5.2]); and three misunderstood the question (e.g., *"I thought it said to choose one other than walking."* [W, 51 y.o., IUIPC: 6.0]).

When asking the participants who did not report a consistent date of their most active day, participants provided a variety of explanations (M.Q28). Similarly, for participants whose self-reported most active day in the past six months and step count differed from the actual values recorded by Fitbit ($n = 77$, Hybrid group), we asked them to explain the discrepancy. The majority ($n = 47$) attributed the difference to forgetfulness, emphasizing the difficulty of recalling such details over a six-month period: *"It's a huge difference. Because I was just guessing."* [W, 31 y.o., IUIPC: 4.6]; *"I did not remember what I walked that long ago, so I chose a more recent day that I remembered."* [W, 48 y.o., IUIPC: 6.0]; *"I am busy a lot of days, but can't really remember day to day how many steps I take."* [W, 33 y.o., IUIPC: 5.5]. The second most common answer, mentioned by $n = 14$ participants, was personal oversight, admitting that they had not checked their Fitbit app before answering. Some were surprised by their actual most active day, whereas others mentioned that they made a close guess. *"I looked back in the app, but I was going quickly and must have missed that day. Looking at the calendar, that was a book shelving day!"* [W, 73 y.o., IUIPC: 5.1].

Unlike the previous question about the most frequent activity type, where many participants raised concerns about Fitbit's accuracy, only $n = 3$ participants mentioned accuracy issues for step count. *"I don't normally walk very much in July because it's too hot. I went on one walk that day, 2.2 miles. I cut grass that day, and the Fitbit counted movement it shouldn't have. I walked much more in September and October."* [W, 46 y.o., IUIPC: 6.0]. This



(a) Perceived Utility (source: M.Q14, M.Q15, and M.Q16).

(b) Perceived Comfort (source: M.Q18, M.Q19, and M.Q20).

Fig. 4. Perceived utility of using a data-driven survey platform and perceived comfort with sharing data for data-driven surveys to answer self-reported data collection questions.

indicates that Fitbit tends to provide more precise step-tracking than activity classification, which is in line with previous literature [23, 33, 39]. Although we did not instruct participants to check (or avoid checking) their Fitbit app when answering, one participant noted that Fitbit's UI limitations made it difficult to report the exact step count: *"The Fitbit data I was looking at was on my Fitbit app on my smartphone in graph form, not showing the actual number; I chose the graph that appeared the highest and then zeroed in on that day to again just get an approximate number. I don't pay for the subscription, so my data access may be limited."* [M, 76 y.o., IUIPC: 7.0]. Lastly, $n = 3$ participants reported misunderstanding the question. *"I think I misread the question and thought you asked about the past month only."* [W, 52 y.o., IUIPC: 5.8].

4.3 User Experience

4.3.1 Utility. Figure 4a shows the perceived utility (M.Q14, M.Q15, and M.Q16) of using a data-driven survey system to answer data-collection questions. The majority (80.7%) of participants found the concept of sharing their data directly, instead of having to manually fill it out, to be at least slightly useful.

Notably, among those who shared their data (i.e., • HYBRID and • DATA-DRIVEN groups), a majority of participants found sharing their data extremely useful—to avoid answering self-reported data collection questions (50% of • HYBRID, and 74% of • DATA-DRIVEN). The Kruskal-Wallis test shows a significant effect of the groups on perceived utility ($\chi^2(2) = 50.4, p < .001$). Post-hoc Mann-Whitney tests confirm significant differences between all group pairs (all $p < .001$), thus indicating that an experience of using a data-driven survey increases its perceived utility.

For the • MANUAL and • HYBRID groups, we analyzed the differences between those who checked their Fitbit data and those who did not. Among the • MANUAL group, those who found it at least slightly useful are 63.9% of those who checked and 71.4% of those who did not. Within the • HYBRID group, those who found it at least slightly useful are 88.1% of those who checked and 82.8% of those who did not. The Kruskal-Wallis test shows a significant difference between the sub-groups (i.e., when broken down by whether participants checked their Fitbit data) ($\chi^2(4) = 52.2, p < .001$). The Mann-Whitney tests further show significant differences (all $p <= .02$) between the following pairs: • MANUAL who checked, and • DATA-DRIVEN and • MANUAL who did not check; • HYBRID who checked, and • DATA-DRIVEN and • MANUAL who did not check; and • HYBRID and • DATA-DRIVEN who did not check. All other pairs are not significant.

4.3.2 Mock-Survey Duration. Table 2 summarizes the median 'mock' survey durations, including the time required to grant access to Fitbit data added for the • HYBRID and • DATA-DRIVEN groups. The participants took

Table 2. Mock-Survey duration for each group (median duration).

Group	Mock-Survey: Factual Duration	Mock-Survey: Perceptual Duration	Total Mock-Survey Duration
• MANUAL Checked Data, Yes	2 min and 21 s	1 min and 57 s	4 min and 51 s
• MANUAL Checked Data, No	1 min and 21 s	1 min and 15 s	2 min and 41 s
• DATA-DRIVEN	1 min and 10 s *	1 min and 48 s	3 min and 15 s

* This is the duration of granting access to their Fitbit accounts on DDS.

9 min and 46 s median time to complete the survey. When looking at the first part of the mock-survey, the Kruskal-Wallis test shows a significant difference ($\chi^2(2) = 33.2, p < .001$). The Mann-Whitney test further shows significant differences between the • MANUAL who checked their data and both the • DATA-DRIVEN ($U = 5413, n_{\text{MANUAL\_CHECK\_YES}} = 72, n_{\text{DATA-DRIVEN}} = 100, p < .001, \text{two–sided}$) and • MANUAL who did not check their data ($U = 1429, n_{\text{MANUAL\_CHECK\_YES}} = 72, n_{\text{MANUAL\_CHECK\_NO}} = 28, p = .004, \text{two–sided}$). There is no significant difference between • DATA-DRIVEN and • MANUAL who did not check their data. This means that the overhead of granting access to data is roughly the same as answering four background information questions about Fitbit data without checking it. Hence, surveys with a greater amount of self-reported data-collection questions may show a greater difference with DDS being notably faster.

4.3.3 Usability (UEQ). Table 3 shows both the hedonic scores (M.Q11) and the pragmatic scores (M.Q11) for each sub-group (i.e., broken down by whether they checked their Fitbit data). The hedonic scores for all groups are all in the range of the 25% worst results relative to the UEQ benchmark data [50]. The Kruskal-Wallis test shows no significant differences between the groups ($\chi^2(2) = 4.4, p = .11$). However, there are significant differences between the sub-groups (i.e., when broken down by whether participants checked their Fitbit data) ($\chi^2(4) = 94.4, p < .001$). The Mann-Whitney tests further show that there are strong significant differences between the • DATA-DRIVEN group and the • MANUAL and • HYBRID sub-groups (all $p < .001$). There are no significant differences between the • MANUAL and • HYBRID sub-groups, and within the • MANUAL and • HYBRID sub-groups.

The pragmatic scores for the • MANUAL and • DATA-DRIVEN groups are excellent relative to the UEQ benchmark data, being in the range of the 10% best results [50]. In contrast, the pragmatic score for the • HYBRID group is good relative to the UEQ benchmark data, where 10% of results are better, 75% of results are worse. The Kruskal-Wallis test shows significant differences between the groups ($\chi^2(2) = 9.7, p = .008$). Post-hoc Mann-Whitney tests further show a significant difference between the • MANUAL and • HYBRID groups ($U = 5156.5, n_{\text{MANUAL}} = 90, n_{\text{HYBRID}} = 92, p = .01, \text{two–sided}$). However, there are no significant differences between • MANUAL and • DATA-DRIVEN ($U = 4293.5, n_{\text{MANUAL}} = 90, n_{\text{DATA-DRIVEN}} = 89, p = 1., \text{two–sided}$), and between • HYBRID and • DATA-DRIVEN ($U = 4922, n_{\text{DATA-DRIVEN}} = 89, n_{\text{HYBRID}} = 92, p = .06, \text{two–sided}$). The Kruskal-Wallis test shows significant differences between the sub-groups ($\chi^2(4) = 10.7, p = .03$). The Mann-Whitney test shows significant differences between the • MANUAL who checked their data and the • HYBRID who did not check their data ($U = 2286, n_{\text{MANUAL}} = 65, n_{\text{HYBRID}} = 54, p = .04, \text{two–sided}$). There are no other significant differences between the sub-groups. In terms of the pragmatic score all groups had a notably good evaluation, with the • MANUAL and • DATA-DRIVEN being very close.

4.3.4 Issues with Granting Access. A modest minority of the participants in the • HYBRID and • DATA-DRIVEN groups experienced issues with granting access to their data (M.Q17), with 17% and 20%, respectively. The most common issue for both groups was not remembering their login information, with 13% and 14%, respectively. A minimal number of the participants in both groups did not understand what they were supposed to do, with 2% and 3%, respectively.

Table 3. UEQ Scores by Score Type ($M \pm SD$). Scores range from -3 to 3.

Group	Hedonic	Pragmatic
● MANUAL Checked, Yes	0.57 ± 1.20	2.00 ± 0.90
● MANUAL Checked, No	0.73 ± 0.78	1.91 ± 1.22
● HYBRID Checked, Yes	0.42 ± 1.03	1.68 ± 1.04
● HYBRID Checked, No	0.26 ± 1.36	1.55 ± 0.95
● DATA-DRIVEN	0.43 ± 1.16	1.96 ± 0.74

A very small fraction of the participants in both groups had "other" problems, with 3% and 7%, for ● HYBRID and ● DATA-DRIVEN, respectively. An issue reported by two participants was the difficulty of switching between Google accounts and remembering which one was linked to Fitbit. *"I had to remember which Google account I had linked to my Fitbit account."* [W, 59 y.o., IUIPC: 6.1]. Another mentioned an issue related to limited access to their Gmail account due to their family account settings. *"My email is connected to a family e-mail, and my Google would not let me connect to my account."* [W, 20 y.o., IUIPC: 6.4]. These issues are mainly due to the specifics of logging into Fitbit accounts after Fitbit was acquired by Google.

Additionally, two participants found reading and verifying the privacy policy to be cumbersome. *"I had to verify and understand what kind of data I am granting access to and how it would be used and kept (possible security/privacy issues)."* [W, 60 y.o., IUIPC: 6.6]. Lastly, two participants reported difficulties logging into Fitbit due to their reliance on saved passwords on their mobile devices. When attempting to log in on a different device, such as a computer, they struggled to recall their credentials. *"Apparently, I forgot the password because it automatically pulls up on my phone I never use the computer to log in."* [W, 49 y.o., IUIPC: 5.1].

4.4 Privacy

4.4.1 Perceived Comfort with Sharing Data. Figure 4b shows the perceived comfort of using a data-driven survey system to answer factual data collection questions (M.Q18, M.Q19, and M.Q20). Overall, just over half of the participants (52%) felt at least slightly comfortable with sharing their data directly instead of having to manually fill it out. Slightly less than half of the participants in the ● MANUAL group (39%) felt at least slightly comfortable with sharing their data. A majority of participants in both the ● HYBRID and ● DATA-DRIVEN groups felt comfortable with sharing their data, with 63% and 54%, respectively. The Kruskal-Wallis test showed significant differences ($\chi^2(2) = 20.3, p < .001$). The Mann-Whitney tests further showed significant differences between the ● MANUAL group and both the ● HYBRID ($U = 3608.5, n_{\text{MANUAL}} = n_{\text{HYBRID}} = 100, p = .002$, two–sided) and ● DATA-DRIVEN ($U = 3315.5, n_{\text{MANUAL}} = n_{\text{DATA-DRIVEN}} = 100, p < .001$, two–sided) groups, where the perceived comfort was lower in the ● MANUAL group. There were no significant differences between the ● HYBRID and ● DATA-DRIVEN groups.

4.4.2 Data-Collection Transparency. A minority of the participants, (31.5%) in both the ● HYBRID and ● DATA-DRIVEN groups, checked the 'privacy transparency table' on DDS,[15] with 30% and 33%, respectively. Overall, among the participants who checked the table, over half of them (66.7%) found that the transparency table and privacy policy was very helpful in convincing them to share their data (M.Q21), with 64.3% and 68.8%, respectively.

4.5 Extra Monetary Compensation

Table 4 shows the median extra monetary compensations that the participants thought that study participants should receive for sharing their data in data-driven surveys, as median values are robust to outliers. A clear majority of the participants (81.7%) believe that study participants should receive extra monetary compensation

[15] This is a feature on the DDS platform that summarizes what data a survey will collect. Figure 5 shows the 'privacy transparency table' shown for this survey.

Table 4. Median Extra Monetary Compensation Required to Share Data by Group and Extra Monetary Compensation Type

Group	Median Absolute Amounts (GBP)	Median Relative Amounts (%)
● MANUAL Checked, Yes	2.8 (n = 44)	22.5% (n = 18)
● MANUAL Checked, No	3 (n = 21)	75% (n = 2)
● HYBRID Checked, Yes	2 (n = 23)	25% (n = 9)
● HYBRID Checked, No	3 (n = 33)	20% (n = 13)
● DATA-DRIVEN	2.5 (n = 63)	20% (n = 19)

when they share their data directly. The proportion of the participants within each group was fairly close, with 85%, 78%, and 82%, for ● MANUAL, ● HYBRID, and ● DATA-DRIVEN groups, respectively. Among those who thought extra monetary compensation should be given, 75.1% of the participants chose to suggest the additional compensation as an *absolute* additional amount, and the rest chose to suggest it as a *relative* amount.

For participating in the survey, participants were paid GBP 1.50 (≈ USD 2). Regarding the actual amounts, participants reported a median *absolute* amount of GBP 2.5 (≈ USD 3.2), and a median *relative* amount of 25% or GBP 0.4 (≈ USD 0.5) when recalculated to absolute terms. Regarding the significance of the absolute amounts, the Kruskal-Wallis test shows a significant difference between the groups for the reported absolute amounts ($\chi^2(2) = 8.1, p = .02$). The Mann-Whitney test only shows a significant difference between the ● MANUAL and ● DATA-DRIVEN groups ($U = 2628, n_{\text{MANUAL}} = n_{\text{DATA-DRIVEN}} = 100, p = .02$, two–sided), wherein the ● MANUAL group reported a median amount of GBP 3 and the ● DATA-DRIVEN group reported a median amount of 2.5. Regarding the significance of the relative amounts, the Kruskal-Wallis test showed no significant differences between both the groups and sub-groups.

5 DISCUSSION

Here, we discuss participants' perspectives on data-driven surveys, highlight seven key insights for researchers, and outline directions for future work.

5.1 Participants' Perspectives on Data-Driven Surveys

Our survey revealed a high valuation of the utility provided by using a data-driven survey system to answer self-reported information-collection questions (Section 4.3.1), regardless of whether they checked their Fitbit app data during the survey. One reason for this could be that manually checking data can be time-consuming, especially when it is information that can be collected automatically. In our survey, having a large proportion of participants who checked their data (Section 4.1) was to be expected, as detailed information about fitness data is not easy for people to recall. Surprisingly, fewer than half of the ● HYBRID participants checked their data. This could be due to them feeling that they did not have to be as precise in their answers because they had already given access to their data.

The analysis of the UEQ (Section 4.3.3) had mixed results. The overall low hedonic scores are somewhat to be expected, as completing surveys is usually not particularly enjoyable. In contrast, the pragmatic scores were very similar between the ● MANUAL and ● DATA-DRIVEN groups, and both were notably higher than the ● HYBRID scores. This could mean that requesting that participants grant access to their data and then asking them to answer factual data collection questions harms their perceptions of the pragmatic qualities of the survey-taking experience. Hence overall, data-driven surveys could improve the pragmatic aspects of survey-taking for participants, especially if the number of factual questions is high. However, they will not have much of an effect on the hedonic aspects.

5.2 Seven Recommendations for Researchers

Based on our findings, we offer seven actionable recommendations for researchers. Some of our recommendations apply broadly to online surveys and some are specific to the use of fitness-tracker data or data-driven surveys. We present the general insights first, followed by the specific ones.

5.2.1 Encourage participants to check their data, but weigh the time cost. Our discrepancy analysis (Section 4.2) showed that the participants who checked their Fitbit data gave more consistent self-reports. Furthermore, even without prompting the participants to check their data, a majority of ● MANUAL participants checked their Fitbit data (Section 4.1). This behavior highlights a low-effort way to improve data accuracy in traditional surveys: simply instruct participants to consult their own data. However, there is a trade-off. Checking data comes with an increased time burden on participants. As shown in Section 4.3.2, those who checked their Fitbit data took significantly longer to complete the mock-survey. Also, not all the requested data (e.g., date of the day with the highest step count) might be easily accessed by the participants (i.e., it is not visible in the the fitness-tracking app).

Since survey length negatively impacts completion rates and data-quality [65, 77], *researchers must carefully balance data accuracy with participant burden. A brief, well-placed prompt to review their data can improve accuracy, but should be used selectively to avoid discouraging completion.*

5.2.2 Match data source to the question—sensor data is not always better. Our qualitative analysis of the discrepancy showed that sometimes relying on data from online accounts can be less reliable than relying on participants' responses. For example, when asking about the most-frequent activity (Section 4.2.1), many participants explained that their Fitbit data was incorrect. This illustrates what Das Swain et al. [20] describe as a 'semantic gap' that is a mismatch between behavioral signals (e.g., logged activities) and participants' subjective perception of their behavior. In our context, Fitbit data may capture *what was done* but not *how participants interpret* their activity, creating discrepancies.

In addition, some inconsistencies are attributable to technical limitations of Fitbit ecosystem. The Fitbit API endpoint, that we used, returns the historically most frequent activity, which may not reflect a participant's current exercise routine.[16] Similarly, regarding weekly exercise-time (Section 4.2.2), the participants who did not check their data were slightly more accurate. This could be due to the Fitbit app UI confusing participants about which metric corresponds to their average weekly exercise-time. Thus, discrepancies arise both from a conceptual misalignment (semantic gap) and technical artifacts (API/UI design), highlighting that sensor data is not always a more reliable ground truth. Hence, *when researchers expect participants to check and self-report data from their online accounts, they should make sure that the requested data can be easily accessed and verified by participants.*

In contrast, for other types of information, data-driven surveys could be more accurate, as the findings from the most-active day date and step count (Section 4.2.3 and Section 4.2.4) show that the further back in time their most active day was, the less accurate participants were in reporting the date. It could be that, for most of our participants, their most active day was not an emotionally charged or important day, which would make it harder to remember [61]. Data-driven surveys can be beneficial in such circumstances, as they can enable presenting participants with additional information, providing more context, to help them to remember why they did something [1]. However, even when fitness trackers provide context, researchers should remain cautious. Previous research shows that clinicians view passive data as valuable but insufficient on its own [71], emphasizing the need to complement it with self-reports for interpretability. To conclude, *researchers should consider the information they collect and how they present it to participants to improve recall and the depth of insights that they provide.*

[16] There is no endpoint that provides the most frequent activity within a specified time range.

5.2.3 Use data-driven surveys to shorten response time and reduce burden. From studying the ‘mock’ survey durations (Section 4.3.2) we found that data-driven surveys could offer a significant time-savings for surveys with extensive background data-collection. The time required to collect data from Fitbit would scale better than the time required to manually find the relevant information in the Fitbit app. For surveys that collect large amounts of data, especially when participants are told to check their data or if they do so of their own volition, using a data-driven survey platform would be beneficial for enabling participants to focus on answering more interesting questions about their data. Therefore, *researchers should consider data-driven surveys a viable solution to reduce survey duration and improve participant experience, especially when factual data are digitally available.*

5.2.4 Offer extra monetary compensation and use data transparency to boost data-sharing. To increase participation in data-driven surveys, researchers should consider two strategies. First, offer extra monetary compensation. In our study (Section 4.5), participants suggested a wide range of reasonable amounts. Hence, *offering extra monetary compensation of approximately the same amount as the baseline survey payment should motivate participants to share their data.*

Second, offer visible transparency into data collection. Our analysis of the data-collection transparency features of DDS (Section 4.4.2) shows that while few participants investigated DDS’s data transparency features, those who did found them to be very helpful in convincing them to share their data. Hence, including such transparency features can be *very* helpful in reassuring and convincing participants to share their data. Therefore, *researchers could also consider linking the privacy policies of systems they use in the survey consent forms or in the study advertisements.*

5.2.5 Disclose data access upfront and over-recruit to anticipate dropouts. Our analyses of the survey responses and completion rates (Table 1) show a notable difference between the ● manual group and the ● hybrid and ● data-driven groups. This is likely due to us not mentioning in the study advertisement that participants would have to grant access to their Fitbit accounts. To save participant time and simplify their decisions about participating in data-driven surveys, *researchers should mention in study advertisements that participants will need to grant access to their data.* This should help to reduce the proportion of participants choosing to withdraw from the study after starting it. To compensate the lower response and completion rates, *researchers should consider recruiting more participants for their data-driven surveys.*

5.2.6 Remind participants to prepare login credentials before starting. A common issue participants encountered with sharing their data was not remembering their login credentials (Section 4.3.4). To improve participant experience and to reduce dropout rates during login steps, *researchers should mention in study advertisements that participants will need to log into their accounts.* Additionally, *researchers could provide instructions for resetting their passwords, before redirecting participants to data-driven survey platforms.*

5.2.7 Anticipate and account for self-selection bias in data-driven surveys. Self-selection bias is a known limitation in survey research, as it can reduce reliability, generalizability, and external validity [8, 47, 106]. In data-driven surveys, this bias can be even more pronounced, due to the more stringent participation criteria. Namely participants would need to (1) be comfortable with granting access to their data, (2) use or have used the desired online platform, and (3) have sufficient digital literacy to grant access to their data. As a result, samples may be skewed toward tech-savvy and/or privacy-tolerant users.

Our results regarding utility (Section 4.3.1) and privacy (Section 4.4.1) indicate that our participants were generally open to data-driven surveys, with the participants in ● hybrid and ● data-driven having higher utility and lower privacy valuations than the ● manual participants. This is most likely due to a self-selection bias because the ● hybrid and ● data-driven participants had to grant access to their data, hence there was a selection for participants with the aforementioned traits. This is further supported by the large difference in survey response and completion rates.

While this bias reduces generalizability it can also help target relevant populations (e.g., users of a particular online service). Nonetheless, *researchers should remember that data-driven surveys can bias samples towards being more tech-savvy and having lower privacy concerns.*

5.3 Limitations

Our study has several limitations. First, the mock survey was fairly short. This limits the maximum time difference we can observe between answering the questions manually and using a data-driven survey system to answer them. Nonetheless, a data-driven survey system should scale better for longer surveys. Second, we did not give participants a choice about granting access to their data. The large differences in completion rates between the ● MANUAL group and the ● HYBRID and ● DATA-DRIVEN groups (see Table 1) suggest that requiring data access reduces participation. Additionally, we did not specify in the Prolific study advertisements for the ● HYBRID and ● DATA-DRIVEN groups that they would need to grant access to their data. This may further reduce the completion rate. Similarly, our response rate estimates show a worst-case scenario because we do not know how many participants from the participant pools had received an invitation to participate. Finally, we only evaluated accessing data through APIs. Methods such as data donation may yield different results. Third, the name of our research institute and lab were visible to participants, which may bias their behavior. However, institutional recognition and trust concerns are common across survey-based studies. Fourth, participants in the ● HYBRID group may not have synchronized their fitness trackers regularly, which can increase discrepancies between the self-reported data and Fitbit-collected data. Fifth, our discrepancy analysis for the most active day (Section 4.2.3 and Section 4.2.4) could be biased towards finding larger discrepancies. Both of the most active day questions (M.Q3 and M.Q4) are much more challenging versions of the types of questions asked in surveys, in contrast to the most frequent activity (M.Q1) and average weekly exercise-time (M.Q2) questions are very similar to questions used in the literature and/or in standard physical activity surveys. When surveys ask about the most-active time within a given time-frame, they usually focus on the most active week day within a week and how much time was spent exercising on that day.

5.4 Future Directions

We first provide future directions to address the limitations of our study, after which we provide overall future directions. First, conducting a survey with a longer 'mock' survey part could help to explore the benefits of using data-driven surveys in greater depth, such as better assessing the time and effort reduction for participants. Second, conducting a survey where participants are given the choice between sharing their data or not could lead to more accurate estimates of the monetary incentives needed to motivate sharing data. To estimate completion rates more accurately in studies that choose to impose granting access to data, study advertisements should clearly communicate needing to grant access to data. To estimate response rates more accurately, researchers can use a platform other than Prolific or they can create a workaround, as currently Prolific does not offer statistics about how many participants it invited. Third, it would be interesting to compare the different ways of conducting data-driven surveys, specifically by conducting comparative studies comparing accessing data via API with data-donation and web-scraping. Several tools exist for data-donation, such as 'Port' that locally processes participants' data, enabling them to select which data they would like to share [9]. Web-scraping, despite gathering large datasets, can be restricted by the terms of use of the platform that one wants to scrape [102]. As each data-driven survey technique has advantages and disadvantages, comparing all of them on a similar survey could provide additional insights about when to use each one.

To go beyond addressing the limitations of our study, several avenues can be explored. First, both self-reported data and data collected from online accounts will have some degree of inaccuracy. Hence, it would be interesting to investigate ways to correct such inaccuracies collaboratively with participants in data-driven surveys. Participants' data could be used to pre-fill/pre-answer the factual parts of the surveys. Doing so would give participants the opportunity to correct any errors that could arise due to the inaccuracy of their online account data, while still maintaining a streamlined survey experience for the data-collection parts. Another approach could be to show the collected data in editable text fields and to ask participants to correct it if necessary. In the context of fitness trackers, this could work well as users tend to be aware of some of the inaccuracies of their devices [6, 79]. Hence, fitness-tracker users should be able to correct inaccuracies in the data collected from their fitness-tracking accounts.   Second, conducting a similar study to develop a more robust statistical model of how relevant factors influence relevant outcomes, such as the perceived utility or the discrepancy between self-reported data and Fitbit data. One possible approach is constructing a model using structured equation modeling (SEM), as there is a complex relationship between demographic factors, internet skills, privacy concerns, data-sharing behavior, and fitness-data valuation [96]. Third, it would be interesting to investigate using data-driven surveys in domains other than fitness tracking; for example, social media and streaming (e.g., audio and video). As privacy concerns and data valuations depend on the data type, user behavior and attitude to data-driven surveys could differ. Furthermore, the discrepancies between self-reported and online account data could be different. In studies on social media, data-driven surveys could access data much closer to the ground truth (i.e., participants' posts) than studies accessing fitness-tracker data. Although there would be limitations, such as inaccessible data (e.g., deleted posts), studying discrepancies may be more reliable and insightful. In studies on streaming habits, data-driven surveys could access data closer to the ground truth (for the streaming service being studied). There may be limitations, such as not knowing what a participant streamed on other platforms. This is similar to the fact that, with Fitbit data, only data related to exercise tracked by a Fitbit wearable is accessible.

6 CONCLUSION

In our study, we have explored the acceptability of data-driven surveys by examining users' willingness to participate, perceived utility and usability, and their privacy concerns. We conducted a mixed-method study by using online surveys with 300 participants. We found that, although participants are generally receptive to the concept of data-driven surveys in fitness-tracking research, they were less likely to participate compared to traditional self-reported surveys. Notably, the discrepancies between participants' self-reported data and their Fitbit data was high, even among those who checked their fitness data. These findings suggest that data-driven surveys could be best suited as a way for collecting self-reported data, before asking follow-up questions. Furthermore, the participants were open to modest extra monetary compensation for participating in data-driven surveys. Overall, our work contributes empirical evidence on the trade-offs and opportunities involved in conducting data-driven surveys for fitness-tracking research and offers actionable guidance for researchers interested in applying these tools in practice. By shedding light on the ways people perceive and engage with data-driven surveys, this work lays a foundation for better understanding and using data-driven survey tools more effectively, in a privacy-conscious way, in future research.

Acknowledgments

This work was partially funded by the Swiss National Science Foundation with Grant #200021_178978 (PrivateLife). We thank Holly Cogliati and Vincent Vandersluis for proofreading this article. We thank Yamane El Zein, Lahari Goswami, Yesle Kim, and Oleksandr Velykoivanenko for participating in the survey cognitive pretests.

References

[1] Deemah Alqahtani, Caroline Jay, and Markel Vigo. 2020. The Role of Uncertainty as a Facilitator to Reflection in Self-Tracking. In *Proc. of the ACM Designing Interactive Systems Conf. (DIS)*. ACM, New York, NY, USA, 1807–1818. doi:10.1145/3357236.3395448

[2] Reza Anaraky, Tahereh Nabizadeh, Bart Knijnenburg, and Marten Risius. 2018. Reducing Default and Framing Effects in Privacy Decision-Making. In *SIGHCI 2018 Proc. Assoc. for Info. Sys. (AIS)*, Vol. 19. AIS, Atlanta, GA, USA, 6.

[3] Chittaranjan Andrade. 2020. The Limitations of Online Surveys. *Indian Journal of Psychological Medicine* 42, 6 (Nov. 2020), 575–576. doi:10.1177/0253717620957496

[4] Lujo Bauer, Lorrie Faith Cranor, Saranga Komanduri, Michelle L. Mazurek, Michael K. Reiter, Manya Sleeper, and Blase Ur. 2013. The Post Anachronism: The Temporal Dimension of Facebook Privacy. In *Proceedings of the 12th ACM Workshop on Workshop on Privacy in the Electronic Society (WPES '13)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/2517840.2517859

[5] Adrian E. Bauman and Justin A. Richards. 2022. Understanding of the Single-Item Physical Activity Question for Population Surveillance. *Journal of Physical Activity and Health* 19, 10 (Oct. 2022), 681–686. doi:10.1123/jpah.2022-0369

[6] Raquel Benbunan-Fich. 2020. User Satisfaction with Wearables. *AIS Transactions on Human-Computer Interaction* 12, 1 (March 2020), 1–27. doi:10.17705/1thci.00126

[7] Brinnae Bent, Ke Wang, Emilia Grzesiak, Chentian Jiang, Yuankai Qi, Yihang Jiang, Peter Cho, Kyle Zingler, Felix Ikponmwosa Ogbeide, Arthur Zhao, Ryan Runge, Ida Sim, and Jessilyn Dunn. 2020. The Digital Biomarker Discovery Pipeline: An Open-Source Software Platform for the Development of Digital Biomarkers Using mHealth and Wearables Data. *Journal of Clinical and Translational Science* 5, 1 (July 2020), e19. doi:10.1017/cts.2020.511

[8] Jelke Bethlehem. 2010. Selection Bias in Web Surveys. *International Statistical Review* 78, 2 (2010), 161–188. doi:10.1111/j.1751-5823.2010.00112.x

[9] Laura Boeschoten, Niek C. de Schipper, Adriënne M. Mendrik, Emiel van der Veen, Bella Struminskaya, Heleen Janssen, and Theo Araujo. 2023. Port: A Software Tool for Digital Data Donation. *Journal of Open Source Software* 8, 90 (Oct. 2023), 5596. doi:10.21105/joss.05596

[10] Norman M. Bradburn, Lance J. Rips, and Steven K. Shevell. 1987. Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science* 236, 4798 (April 1987), 157–161. doi:10.1126/science.3563494

[11] Virginia Braun, Victoria Clarke, Elicia Boulton, Louise Davey, and Charlotte McEvoy. 2021. The Online Survey as a Qualitative Research Tool. *International Journal of Social Research Methodology* 24, 6 (Nov. 2021), 641–654. doi:10.1080/13645579.2020.1805550

[12] Axel Buchner, Edgar Erdfelder, Franz Faul, and Albert-Georg Lang. 2023. G *Power 3.1 Manual.

[13] Aaron Carpenter. 2024. How to Increase Online Survey Response Rates.

[14] Thomas P. Carpenter, Ruth Pogacar, Chris Pullig, Michal Kouril, Stephen Aguilar, Jordan LaBouff, Naomi Isenberg, and Alek Chakroff. 2019. Survey-Software Implicit Association Tests: A Methodological and Empirical Analysis. *Behavior Research Methods* 51, 5 (Oct. 2019), 2194–2208. doi:10.3758/s13428-019-01293-3

[15] Thijs C. Carrière, Laura Boeschoten, Bella Struminskaya, Heleen L. Janssen, Niek C. de Schipper, and Theo Araujo. 2024. Best Practices for Studies Using Digital Data Donation. *Quality & Quantity* 59, 1 (Oct. 2024), 389–412. doi:10.1007/s11135-024-01983-x

[16] Irene Celino and Gloria Re Calegari. 2020. Submitting Surveys via a Conversational Interface: An Evaluation of User Acceptance and Approach Effectiveness. *International Journal of Human-Computer Studies* 139 (July 2020), 102410. doi:10.1016/j.ijhcs.2020.102410

[17] Pilsik Choi and Keith S. Coulter. 2012. It's Not All Relative: The Effects of Mental and Physical Positioning of Comparative Prices on Absolute versus Relative Discount Assessment. *Journal of Retailing* 88, 4 (Dec. 2012), 512–527. doi:10.1016/j.jretai.2012.04.001

[18] Chia-Fang Chung, Nanna Gorm, Irina A. Shklovski, and Sean Munson. 2017. Finding the Right Fit: Understanding Health Tracking in Workplace Wellness Programs. In *CHI*. ACM, New York, NY, USA, 4875–4886. doi:10.1145/3025453.3025510

[19] Jiska Classen, Daniel Wegemer, Paul Patras, Tom Spink, and Matthias Hollick. 2018. Anatomy of a Vulnerable Fitness Tracking System: Dissecting the Fitbit Cloud, App, and Firmware. *IMWUT* 2, 1 (March 2018), 5:1–5:24. doi:10.1145/3191737

[20] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M. Mattingly, Gregory D. Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3491102.3502037

[21] Martyn Denscombe. 2010. *The Good Research Guide: For Small-Scale Social Research Projects* (4th edition ed.). Open University Press, Maidenhead.

[22] Don A. Dillman, Jolene D. Smyth, and Leah Melani Christian. 2014. *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method, 4th Edition | Wiley* (4th edition ed.). Wiley, Hoboken.

[23] Diana Dorn, Jessica Gorzelitz, Ronald Gangnon, David Bell, Kelli Koltyn, and Lisa Cadmus-Bertram. 2019. Automatic Identification of Physical Activity Type and Duration by Wearable Activity Trackers: A Validation Study. *JMIR mHealth and uHealth* 7, 5 (May 2019), e13547. doi:10.2196/13547

[24] Nickolas Dreher, Edward Kenji Hadeler, Sheri J. Hartman, Emily C. Wong, Irene Acerbi, Hope S. Rugo, Melanie Catherine Majure, Amy Jo Chien, Laura J. Esserman, and Michelle E. Melisko. 2019. Fitbit Usage in Patients With Breast Cancer Undergoing Chemotherapy. *Clinical Breast Cancer* 19, 6 (Dec. 2019), 443–449.e1. doi:10.1016/j.clbc.2019.05.005

[25] Nico Ebert, Björn Scheppler, Kurt Alexander Ackermann, and Tim Geppert. 2023. QButterfly: Lightweight Survey Extension for Online User Interaction Studies for Non-Tech-Savvy Researchers. In *Proc. of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, Hamburg Germany, 1–8. doi:10.1145/3544548.3580780

[26] Daniel A. Epstein, Monica Caraway, Chuck Johnston, An Ping, James Fogarty, and Sean A. Munson. 2016. Beyond Abandonment to Next Steps: Understanding and Designing for Life after Personal Informatics Tool Use. In *CHI*. ACM, San Jose California USA, 1109–1113. doi:10.1145/2858036.2858045

[27] Daniel A. Epstein, Jennifer H. Kang, Laura R. Pina, James Fogarty, and Sean A. Munson. 2016. Reconsidering the Device in the Drawer: Lapses as a Design Opportunity in Personal Informatics. In *Proc. of the Conf. on Ubiquitous Computing (UbiComp)*. Association for Computing Machinery, Heidelberg, Germany, 829–840. doi:10.1145/2971648.2971656

[28] Joel R. Evans and Anil Mathur. 2005. The Value of Online Surveys. *Internet Research* 15, 2 (Jan. 2005), 195–219. doi:10.1108/10662240510590360

[29] Barry Fass-Holmes. 2022. Survey Fatigue–What Is Its Role in Undergraduates' Survey Participation and Response Rates? *Journal of Interdisciplinary Studies in Education* 11, 1 (2022), 56–73.

[30] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical Power Analyses Using G*Power 3.1: Tests for Correlation and Regression Analyses. *Behavior Research Methods* 41, 4 (Nov. 2009), 1149–1160. doi:10.3758/BRM.41.4.1149

[31] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2020. G*Power.

[32] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behavior Research Methods* 39, 2 (May 2007), 175–191. doi:10.3758/BF03193146

[33] Lynne M. Feehan, Jasmina Geldman, Eric C. Sayre, Chance Park, Allison M. Ezzat, Ju Young Yoo, Clayon B. Hamilton, and Linda C. Li. 2018. Accuracy of Fitbit Devices: Systematic Review and Narrative Syntheses of Quantitative Data. *JMIR mHealth and uHealth* 6, 8 (Aug. 2018). doi:10.2196/10527

[34] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. 2015. AWARE: Mobile Context Instrumentation Framework. *Frontiers in ICT* 2 (April 2015). doi:10.3389/fict.2015.00006

[35] Janet Finlayson, Angela Turner, and Malcolm H. Granat. 2011. Measuring the Actual Levels and Patterns of Physical Activity/Inactivity of Adults with Intellectual Disabilities. *Journal of Applied Research in Intellectual Disabilities* 24, 6 (Nov. 2011), 508–517. doi:10.1111/j.1468-3148.2011.00633.x

[36] Sandra Gabriele and Sonia Chiasson. 2020. Understanding Fitness Tracker Users' Security and Privacy Knowledge, Attitudes and Behaviours. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376651

[37] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding Malicious Behavior in Crowdsourcing Platforms: The Case of Online Surveys. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 1631–1640. doi:10.1145/2702123.2702443

[38] Pablo Galan-Lopez and Francis Ries. 2019. Motives for Exercising and Associations with Body Composition in Icelandic Adolescents. *Sports* 7, 6 (June 2019), 149. doi:10.3390/sports7060149

[39] Federico Germini, Noella Noronha, Victoria Borg Debono, Binu Abraham Philip, Drashti Pete, Tamara Navarro, Arun Keepanasseril, Sameer Parpia, Kerstin de Wit, and Alfonso Iorio. 2022. Accuracy and Acceptability of Wrist-Wearable Activity-Tracking Devices: Systematic Review of the Literature. *Journal of Medical Internet Research* 24, 1 (Jan. 2022), e30791. doi:10.2196/30791

[40] Google. 2024. How Do I Track My Activity with My Fitbit Device? - Fitbit Help Center.

[41] Anja S. Göritz. 2014. Determinants of the Starting Rate and the Completion Rate in Online Panel Studies. In *Online Panel Research* (1 ed.), Mario Callegaro, Reg Baker, Jelke Bethlehem, Anja S. Göritz, Jon A. Krosnick, and Paul J. Lavrakas (Eds.). Wiley, West Sussex, United Kingdom, 154–170. doi:10.1002/9781118763520.ch7

[42] Ashley K Griggs, Marcus E Berzofsky, Bonnie E Shook-Sa, Christine H Lindquist, Kimberly P Enders, Christopher P Krebs, Michael Planty, and Lynn Langton. 2018. The Impact of Greeting Personalization on Prevalence Estimates in a Survey of Sexual Assault Victimization. *Public Opinion Quarterly* 82, 2 (June 2018), 366–378. doi:10.1093/poq/nfy019

[43] Pamela Grimm. 2010. Social Desirability Bias. In *Wiley International Encyclopedia of Marketing* (1st ed.), Jagdish Sheth and Naresh K Malhotra (Eds.). Vol. 2. John Wiley & Sons, Ltd, West Sussex, United Kingdom. doi:10.1002/9781444316568.wiem02057

[44] Thomas Groß. 2021. Validity and Reliability of the Scale Internet Users' Information Privacy Concerns (IUIPC). *Proceedings on Privacy Enhancing Technologies* 2021, 2 (April 2021), 235–258. doi:10.2478/popets-2021-0026

[45] Lee Harrison, Mark A. Brennan, and Alan M. Levine. 2000. Physical Activity Patterns and Body Mass Index Scores among Military Service Members. *American Journal of Health Promotion* 15, 2 (Nov. 2000), 77–80. doi:10.4278/0890-1171-15.2.77

[46] Valerie Hase, Jef Ausloos, Laura Boeschoten, Nico Pfiffner, Heleen Janssen, Theo Araujo, Thijs Carrière, Claes de Vreese, Jörg Haßler, Felicia Loecherbach, Zoltán Kmetty, Judith Möller, Jakob Ohme, Elisabeth Schmidbauer, Bella Struminskaya, Damian Trilling, Kasper Welbers, and Mario Haim. 2024. Fulfilling Data Access Obligations: How Could (and Should) Platforms Facilitate Data Donation Studies? *Internet Policy Review* 13, 3 (Sept. 2024).

[47] James J. Heckman. 1979. Sample Selection Bias as a Specification Error. *Econometrica* 47, 1 (Jan. 1979), 153. jstor:1912352 doi:10.2307/1912352

[48] Dirk Heerwegh, Tim Vanhove, Koen Matthijs, and Geert Loosveldt. 2005. The Effect of Personalization on Response Rates and Data Quality in Web Surveys. *International Journal of Social Research Methodology* 8, 2 (April 2005), 85–99. doi:10.1080/1364557042000203107

[49] André Henriksen, Martin Haugen Mikalsen, Ashenafi Zebene Woldaregay, Miroslav Muzny, Gunnar Hartvigsen, Laila Arnesdatter Hopstock, and Sameline Grimsgaard. 2018. Using Fitness Trackers and Smartwatches to Measure Physical Activity in Research: Analysis of Consumer Wrist-Worn Wearables. *Journal of Medical Internet Research* 20, 3 (March 2018), e9157. doi:10.2196/jmir.9157

[50] Andreas Hinderks, Martin Schrepp, and Jörg Thomaschewski. 2024. User Experience Questionnaire.

[51] Evana T. Hsiao and Robert E. Thayer. 1998. Exercising for Mood Regulation: The Importance of Experience. *Personality and Individual Differences* 24, 6 (June 1998), 829–836. doi:10.1016/S0191-8869(98)00013-0

[52] Kevin Huguenin, Igor Bilogrevic, Joana Soares Machado, Stefan Mihaila, Reza Shokri, Italo Dacosta, and Jean-Pierre Hubaux. 2018. A Predictive Model for User Motivation and Utility Implications of Privacy-Protection Mechanisms in Location Check-Ins. *IEEE Transactions on Mobile Computing* 17, 4 (April 2018), 760–774. doi:10.1109/TMC.2017.2741958

[53] Rune Møberg Jacobsen, Samuel Rhys Cox, Carla F. Griggio, and Niels van Berkel. 2025. Chatbots for Data Collection in Surveys: A Comparison of Four Theory-Based Interview Probes. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, Yokohama, Japan, 1–21. doi:10.1145/3706598.3714128

[54] Maritza Johnson, Serge Egelman, and Steven M. Bellovin. 2012. Facebook and Privacy: It's Complicated. In *Proceedings of the Symposium on Usable Privacy and Security*. Association for Computing Machinery, Washington, D.C., 1–15. doi:10.1145/2335356.2335369

[55] Soomin Kim, Joonhwan Lee, and Gahgene Gweon. 2019. Comparing Data from Chatbot and Web Surveys: Effects of Platform and Conversational Style on Survey Response Quality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–12. doi:10.1145/3290605.3300316

[56] Yong-Wook Kim, Jinyoung Han, Kyungtae Jang, Minsam Ko, Jaewoo Park, Seungyup Lim, and Jin-Young Lee. 2022. The Connection to the Public's Preferred Sports Analysis and Physical Education Curriculum. *PLOS ONE* 17, 3 (March 2022), e0264032. doi:10.1371/journal.pone.0264032

[57] Robert J. Kirkby, Gregory S. Kolt, Kathleen Habel, and Jeremy Adams. 1999. Exercise in Older Women: Motives for Participation. *Australian Psychologist* 34, 2 (July 1999), 122–127. doi:10.1080/00050069908257440

[58] Frauke Kreuter, Georg-Christoph Haas, Florian Keusch, Sebastian Bähr, and Mark Trappmann. 2020. Collecting Survey and Smartphone Sensor Data With an App: Opportunities and Challenges Around Privacy and Informed Consent. *Social Science Computer Review* 38, 5 (Oct. 2020), 533–549. doi:10.1177/0894439318816389

[59] Simon Kühne and Martin Kroh. 2018. Personalized Feedback in Web Surveys: Does It Affect Respondents' Motivation and Data Quality? *Social Science Computer Review* 36, 6 (Dec. 2018), 744–755. doi:10.1177/0894439316673604

[60] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. Chapter 10 - Usability Testing. In *Research Methods in Human Computer Interaction (Second Edition)*, Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser (Eds.). Morgan Kaufmann, Boston, 263–298.

[61] Hongmi Lee and Janice Chen. 2022. Predicting Memory from the Network Structure of Naturalistic Events. *Nature Communications* 13, 1 (July 2022), 4235. doi:10.1038/s41467-022-31965-2

[62] Hyunsoo Lee, Soowon Kang, and Uichin Lee. 2022. Understanding Privacy Risks and Perceived Benefits in Open Dataset Collection for Mobile Affective Computing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (July 2022), 61:1–61:26. doi:10.1145/3534623

[63] Ian Li, Anind K. Dey, and Jodi Forlizzi. 2011. Understanding My Data, Myself: Supporting Self-Reflection with Ubicomp Technologies. In *Proceedings of the 13th International Conference on Ubiquitous Computing (UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 405–414. doi:10.1145/2030112.2030166

[64] Brian Y. Lim, Judy Kay, and Weilong Liu. 2019. How Does a Nation Walk? Interpreting Large-Scale Step Count Activity with Weekly Streak Patterns. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (June 2019), 57:1–57:46. doi:10.1145/3328928

[65] Mingnan Liu and Laura Wronski. 2018. Examining Completion Rates in Web Surveys via Over 25,000 Real-World Surveys. *Social Science Computer Review* 36, 1 (Feb. 2018), 116–124. doi:10.1177/0894439317695581

[66] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. 2004. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research* 15, 4 (2004), 336–355. jstor:23015787

[67] Tenga Matsuura, Ayako A. Hasegawa, Mitsuaki Akiyama, and Tatsuya Mori. 2021. Careless Participants Are Essential for Our Phishing Study: Understanding the Impact of Screening Methods. In *Proceedings of the 2021 European Symposium on Usable Security (EuroUSEC '21)*. Association for Computing Machinery, New York, NY, USA, 36–47. doi:10.1145/3481357.3481515

[68] Vikas Menon and Aparna Muraleedharan. 2020. Internet-Based Surveys: Relevance, Methodological Considerations and Troubleshooting Strategies. *General Psychiatry* 33, 5 (Aug. 2020), e100264. doi:10.1136/gpsych-2020-100264

[69] Andras Molnar. 2019. SMARTRIQS: A Simple Method Allowing Real-Time Respondent Interaction in Qualtrics Surveys. *Journal of Behavioral and Experimental Finance* 22 (June 2019), 161–169. doi:10.1016/j.jbef.2019.03.005

[70] Ashwini Nagappan, Adriana Krasniansky, and Madelyn Knowles. 2024. Patterns of Ownership and Usage of Wearable Devices in the United States, 2020-2022: Survey Study. *Journal of Medical Internet Research* 26, 1 (July 2024), e56504. doi:10.2196/56504

[71] Jodie Nghiem, Daniel A. Adler, Deborah Estrin, Cecilia Livesey, and Tanzeem Choudhury. 2023. Understanding Mental Health Clinicians' Perceptions and Concerns Regarding Using Passive Patient-Generated Health Data for Clinical Decision-Making: Qualitative Semistructured Interview Study. *JMIR Formative Research* 7, 1 (Aug. 2023), e47380. doi:10.2196/47380

[72] Yuuki Nishiyama, Denzil Ferreira, Yusaku Eigen, Wataru Sasaki, Tadashi Okoshi, Jin Nakazawa, Anind K. Dey, and Kaoru Sezaki. 2020. IOS Crowd–Sensing Won't Hurt a Bit!: AWARE Framework and Sustainable Study Guideline for iOS Platform. In *Distributed, Ambient and Pervasive Interactions*, Norbert Streitz and Shin'ichi Konomi (Eds.). Springer International Publishing, Cham, 223–243. doi:10.1007/978-3-030-50344-4_17

[73] Yuuki Nishiyama, Denzil Ferreira, Wataru Sasaki, Tadashi Okoshi, Jin Nakazawa, Anind K. Dey, and Kaoru Sezaki. 2020. Using iOS for Inconspicuous Data Collection: A Real-World Assessment. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp/ISWC '20 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 261–266. doi:10.1145/3410530.3414369

[74] Jason Orlosky, Onyeka Ezenwoye, Heather Yates, and Gina Besenyi. 2019. A Look at the Security and Privacy of Fitbit as a Health Activity Tracker. In *Proceedings of the 2019 ACM Southeast Conference (ACM SE '19)*. Association for Computing Machinery, Kennesaw, GA, USA, 241–244. doi:10.1145/3299815.3314468

[75] Xinru Page, Paritosh Bahirat, Muhammad I. Safi, Bart P. Knijnenburg, and Pamela Wisniewski. 2018. The Internet of What? Understanding Differences in Perceptions and Adoption for the Internet of Things. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (Dec. 2018), 183:1–183:22. doi:10.1145/3287061

[76] Stefan Palan and Christian Schitter. 2018. Prolific.Ac—A Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* 17 (March 2018), 22–27. doi:10.1016/j.jbef.2017.12.004

[77] Stephen R. Porter, Michael E. Whitcomb, and William H. Weitzer. 2004. Multiple Surveys of Students and Survey Fatigue. *New Directions for Institutional Research* 2004, 121 (2004), 63–73. doi:10.1002/ir.101

[78] Yatharth Ranjan, Zulqarnain Rashid, Callum Stewart, Pauline Conde, Mark Begale, Denny Verbeeck, Sebastian Boettcher, The Hyve, Richard Dobson, Amos Folarin, and The RADAR-CNS Consortium. 2019. RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. *JMIR mHealth and uHealth* 7, 8 (Aug. 2019), e11734. doi:10.2196/11734

[79] Amon Rapp and Lia Tirabeni. 2018. Personal Informatics for Sport: Meaning, Body, and Social Relations in Amateur and Elite Athletes. *ACM Transactions on Computer-Human Interaction* 25, 3 (June 2018), 16:1–16:30. doi:10.1145/3196829

[80] Jungwook Rhim, Minji Kwak, Yeaeun Gong, and Gahgene Gweon. 2022. Application of Humanization to Survey Chatbots: Change in Chatbot Perception, Interaction Experience, and Survey Data Quality. *Computers in Human Behavior* 126 (Jan. 2022), 107034. doi:10.1016/j.chb.2021.107034

[81] Ritesh Saini, Raghunath Singh Rao, and Ashwani Monga. 2010. Is That Deal Worth My Time? The Interactive Effect of Relative and Referent Thinking on Willingness to Seek a Bargain. *Journal of Marketing* 74, 1 (Jan. 2010), 34–48. doi:10.1509/jmkg.74.1.34

[82] Johnny Saldana. 2021. *The Coding Manual for Qualitative Researchers* (4e [fourth editiion] ed.). SAGE Publishing Inc, Thousand Oaks, California.

[83] Kavous Salehzadeh Niksirat, Lahari Goswami, Pooja S. B. Rao, James Tyler, Alessandro Silacci, Sadiq Aliyu, Annika Aebli, Chat Wacharamanotham, and Mauro Cherubini. 2023. Changes in Research Ethics, Openness, and Transparency in Empirical Studies between CHI 2017 and CHI 2022. In *Proc. of the CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, 1–23. doi:10.1145/3544548.3580848

[84] Kavous Salehzadeh Niksirat, Lev Velykoivanenko, Noé Zufferey, Mauro Cherubini, Kévin Huguenin, and Mathias Humbert. 2024. Wearable Activity Trackers: A Survey on Utility, Privacy, and Security. *Comput. Surveys* 56, 7 (July 2024), 1–40. doi:10.1145/3645091

[85] Roberta Sammut, Odette Griscti, and Ian J. Norman. 2021. Strategies to Improve Response Rates to Web Surveys: A Literature Review. *International Journal of Nursing Studies* 123 (Nov. 2021), 104058. doi:10.1016/j.ijnurstu.2021.104058

[86] Andrea Schankin, Matthias Budde, Till Riedel, and Michael Beigl. 2022. Psychometric Properties of the User Experience Questionnaire (UEQ). In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3491102.3502098

[87] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2014. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience*, Aaron Marcus (Ed.). Springer International Publishing, Cham, 383–392. doi:10.1007/978-3-319-07668-3_37

[88] Martin Schrepp, Jessica Kollmorgen, and Jörg Thomaschewski. 2023. A Comparison of SUS, UMUX-LITE, and UEQ-S. *Journal of User Experience* 18, 2 (2023), 86–104.

[89] Grace Shin, Mohammad Hossein Jarrahi, Yu Fei, Amir Karami, Nicci Gafinowitz, Ahjung Byun, and Xiaopeng Lu. 2019. Wearable Activity Trackers, Accuracy, Adoption, Acceptance and Health Impact: A Systematic Literature Review. *Journal of Biomedical Informatics* 93 (May 2019), 103153. doi:10.1016/j.jbi.2019.103153

[90] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. 2019. How to Do Better with Gender on Surveys: A Guide for HCI Researchers. *Interactions* 26, 4 (June 2019), 62–65. doi:10.1145/3338283

[91] Jan-Benedict E.M. Steenkamp, Martijn G. De Jong, and Hans Baumgartner. 2010. Socially Desirable Response Tendencies in Survey Research. *Journal of Marketing Research* 47, 2 (April 2010), 199–214. doi:10.1509/jmkr.47.2.199

[92] Niels van Berkel, Simon D'Alfonso, Rio Kurnia Susanto, Denzil Ferreira, and Vassilis Kostakos. 2023. AWARE-Light: A Smartphone Tool for Experience Sampling and Digital Phenotyping. *Personal and Ubiquitous Computing* 27, 2 (April 2023), 435–445. doi:10.1007/s00779-022-01697-7

[93] András Vargha and Harold D. Delaney. 1998. The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics* 23, 2 (June 1998), 170–192. doi:10.3102/10769986023002170

[94] Lev Velykoivanenko, Kavous Salehzadeh Niksirat, Noé Zufferey, Mathias Humbert, Kévin Huguenin, and Mauro Cherubini. 2022. Are Those Steps Worth Your Privacy? Fitness-Tracker Users' Perceptions of Privacy and Utility. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5, 4 (Dec. 2022), 181:1–181:41. doi:10.1145/3494960

[95] Lev Velykoivanenko, Kavous Salehzadeh Niksirat, Stefan Teofanovic, Bertil Chapuis, Michelle L. Mazurek, and Kévin Huguenin. 2024. Designing a Data-Driven Survey System: Leveraging Participants' Online Data to Personalize Surveys. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–22. doi:10.1145/3613904.3642572

[96] Jessica Vitak, Yuting Liao, Priya Kumar, Michael Zimmer, and Katherine Kritikos. 2018. Privacy Attitudes and Data Valuation Among Fitness Tracker Users. In *Transforming Digital Worlds (Lecture Notes in Computer Science)*, Gobinda Chowdhury, Julie McLeod, Val Gillet, and Peter Willett (Eds.). Springer International Publishing, Cham, 229–239. doi:10.1007/978-3-319-78105-1_27

[97] Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging Large Language Models to Power Chatbots for Collecting User Self-Reported Data. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 87:1–87:35. doi:10.1145/3637364

[98] Miranda Wei, Madison Stamos, Sophie Veys, Nathan Reitinger, Justin Goodman, Margot Herman, Dorota Filipczuk, Ben Weinshel, Michelle L. Mazurek, and Blase Ur. 2020. What Twitter Knows: Characterizing Ad Targeting Practices, User Perceptions, and Ad Explanations Through Users' Own Twitter Data. In *Proceedings of the 29th USENIX Conference on Security Symposium (SEC'20)*. USENIX Association, USA, 145–162.

[99] James Wen and Ashley Colley. 2022. Hybrid Online Survey System with Real-Time Moderator Chat. In *Proceedings of the 21st International Conference on Mobile and Ubiquitous Multimedia (MUM '22)*. Association for Computing Machinery, New York, NY, USA, 257–258. doi:10.1145/3568444.3570593

[100] Meng-Jia Wu, Kelly Zhao, and Francisca Fils-Aime. 2022. Response Rates of Online Surveys in Published Research: A Meta-Analysis. *Computers in Human Behavior Reports* 7 (Aug. 2022), 100206. doi:10.1016/j.chbr.2022.100206

[101] Ziang Xiao, Michelle X. Zhou, Q. Vera Liao, Gloria Mark, Changyan Chi, Wenxi Chen, and Huahai Yang. 2020. Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM Transactions on Computer-Human Interaction* 27, 3 (June 2020), 15:1–15:37. doi:10.1145/3381804

[102] Cai Yang, Sepehr Mousavi, Abhisek Dash, Krishna P. Gummadi, and Ingmar Weber. 2024. Coupling GDPR Data Donation and Crowdsourced User Survey: A Case Study on TikTok Addiction. In *Companion Proceedings of the 16th ACM Web Science Conference*. ACM, Stuttgart Germany, 51–52. doi:10.1145/3630744.3659830

[103] Koji Yatani. 2014. Statistical Methods for HCI Research.

[104] Florian Zandt. 2023. The Most Popular Sports & Activities in the U.S.

[105] Brahim Zarouali, Theo Araujo, Jakob Ohme, and Claes de Vreese. 2023. Comparing Chatbots and Online Surveys for (Longitudinal) Data Collection: An Investigation of Response Characteristics, Data Quality, and User Evaluation. *Communication Methods and Measures* 0, 0 (Jan. 2023), 1–20. doi:10.1080/19312458.2022.2156489

[106] Qian Zhu, Leo Yu-Ho Lo, Meng Xia, Zixin Chen, and Xiaojuan Ma. 2022. Bias-Aware Design for Informed Decisions: Raising Awareness of Self-Selection Bias in User Ratings and Reviews. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022), 496:1–496:31. doi:10.1145/3555597

[107] Olena Zimba and Armen Yuri Gasparyan. 2023. Designing, Conducting, and Reporting Survey Studies: A Primer for Researchers. *Journal of Korean Medical Science* 38, 48 (Nov. 2023), 11. doi:10.3346/jkms.2023.38.e403

[108] Michael Zimmer, Priya Kumar, Jessica Vitak, Yuting Liao, and Katie Chamberlain Kritikos. 2020. 'There's Nothing Really They Can Do with This Information': Unpacking How Users Manage Privacy Boundaries for Personal Fitness Information. *Information, Communication & Society* 23, 7 (June 2020), 1020–1037. doi:10.1080/1369118X.2018.1543442

[109] Noé Zufferey, Mathias Humbert, Romain Tavenard, and Kévin Huguenin. 2023. Watch Your Watch: Inferring Personality Traits from Wearable Activity Trackers. In *Proceedings of the 32nd USENIX Conference on Security Symposium (SEC '23)*. USENIX Association, Anaheim, CA, USA, 193–210.

[110] Noé Zufferey, Kavous Salehzadeh Niksirat, Mathias Humbert, and Kévin Huguenin. 2023. "Revoked Just Now!" Users' Behaviors Toward Fitness-Data Sharing with Third-Party Applications. *Proceedings on Privacy Enhancing Technologies* (2023).

[111] Noé Zufferey, Kavous Salehzadeh Niksirat, Mathias Humbert, and Kévin Huguenin. 2024. Our Data, Our Solutions: A Participatory Approach for Enhancing Privacy in Wearable Activity Tracker Third-Party Apps. *Proceedings on Privacy Enhancing Technologies* 2024, 4 (2024), 734–754. doi:10.56553/popets-2024-0139

A DETAILED SURVEY DESIGN

Here, we present both the consent survey (Section B.1) and the main survey (Section B.2) as a single streamlined survey, as a participant would experience it. Sections and questions prefixed with a "C." refer to the consent survey, and those prefixed with an "M." refer to the main survey. The survey comprises two primary parts: the first was the mock survey (i.e., M.Part 1 and M.Part 2), and the second was about survey taking experiences (i.e., from M.Part 3 to M.Part 8).

Part 1 This part (C.Part 1) asked the participants two questions to confirm their Prolific ID (C.Q1) and to consent to participate in the survey (C.Q2).

Part 2 This part (M.Part 1) was the **factual** part of the 'mock' survey; it comprised four questions (M.Q1-M.Q4) for the collection of factual information about the physical activity-related data recorded in the participants' Fitbit accounts and was answered by only the • MANUAL and • HYBRID groups. To create these questions, we looked at what information is collected in the literature [5, 18, 35, 49, 56], could be retrieved from the Fitbit app UI (as we anticipated many participants checking their data), and could be computed from the data accessed via the Fitbit API. To minimize discrepancies in the participants' answers, we phrased all our questions to ask them to report only information about when they were wearing their Fitbit device. Specifically, the questions were:

- Most frequent activity [M.Q1, based on [40, 56, 104]]: "On a typical week, **which activity** (other than **Walk**) is **most often** detected automatically or logged manually in your Fitbit account?"
- Average weekly exercise-time [M.Q2, based on [5]]: "How many **minutes** do you usually spend **exercising** in total over an entire **typical week** while **wearing** your **Fitbit**?"
- Most active day—date [M.Q3, based on [35]]: "Approximately on **which day** over the **last six months** do you think you had your **highest step count** while wearing your Fitbit (yyyy-mm-dd)?"
- Most active day—steps [M.Q4, based on [35]]: "Approximately how many **steps** do you think you took on **that day** (${most active day—date}[M.Q3])?"

Part 3 This part (M.Part 2) was the **perceptual part** of the 'mock' survey. There were three questions for the • MANUAL and • HYBRID groups (M.Q5-M.Q7) and three questions for the • DATA-DRIVEN group (M.Q8-M.Q10); we collected perceptual information related to each question asked in the first part of the mock survey. The questions were meant to elaborate on the participants' answers to the factual part of the mock survey. For example, "You reported that you do **${exercise time} minutes** of **exercise per week** while wearing your Fitbit. What are your **main reason(s)** or **motivation(s)** for **doing this amount** of exercise?" The • DATA-DRIVEN group's questions were "answered" using their Fitbit data. The options for the follow-up question about motivation for exercising (M.Q6) were based on prior literature [38, 45, 51, 57].

Part 4 This part (M.Part 3) comprised three questions for evaluating the usability of the mock survey taking experience (see RQ1). M.Q11 was the short User Experience Questionnaire (UEQ) [87, 88]. The short UEQ has been shown to be a reliable way to assess the hedonic and pragmatic aspects of user experiences (UX). The hedonic scale measures the level of enjoyment the users felt using a system [86, 88]. In this context, this refers to the level of enjoyment the participants had in completing the mock survey. The pragmatic scale measures the level of support the users felt in efficiently and effectively accomplishing a set task [86, 88]. In this context, this refers to the level of efficiency and effectiveness the participants felt about completing the mock survey. We chose to use the UEQ, as it provides a reliable measure of users' experiences [86, 88], while measuring both the pragmatic and hedonic aspects. In contrast, other popular usability questionnaires such as the System Usability Score (SUS) and the Usability Metric for User Experience-Lite Version UMUX-LITE measure only the pragmatic quality [88]. M.Q12 and M.Q13 were the same question with slightly different phrasing, for the • MANUAL and • HYBRID groups; they asked whether the participants checked their Fitbit data while answering the mock survey questions. We hypothesized that checking Fitbit data increases the time and effort required to complete a survey, hence it could influence participants' perceptions of data-driven surveys in later questions.

Part 5 This part (M.Part 4) comprised four questions for evaluating the perceived utility and usability of data-driven surveys (see RQ1 and RQ1). Each group was asked how useful using a data-driven survey to answer self-reported data collection questions would be (M.Q14, M.Q15, and M.Q16). We compared the perceived utility of a hypothetical usage scenario (● MANUAL) with a realized one (● DATA-DRIVEN), and one where it could have been done (● HYBRID). Then the ● HYBRID and ● DATA-DRIVEN groups were asked which issues they experienced (if any) with granting access to their Fitbit data (M.Q17). As data-driven surveys are not commonly used, it is important to identify common issues that participants are likely to experience.

Part 6 This part (M.Part 5) comprises five questions for evaluating the privacy aspects of taking data-driven surveys (see RQ1). Each group was asked how comfortable they would be with granting access to their Fitbit data for a data-driven survey to answer self-reported data collection questions (M.Q18, M.Q19, and M.Q20). We evaluated the level of privacy concern with regard to granting access to their Fitbit data for a data-driven survey. Then, the participants in the ● HYBRID and ● DATA-DRIVEN groups, which opened the 'collected data table' on DDS, were asked how much the privacy and transparency features of DDS helped convince them to grant access to their data (M.Q21). We learned to what extent the participants, who chose to investigate the privacy and transparency aspects of our data-driven survey, would be convinced by them. This is important as improved transparency and data protection could convince more participants to participate in data-driven surveys.

Part 7 This part (M.Part 6) comprises five questions for evaluating the amount of additional monetary incentive that would be needed to motivate participants to share their data (see RQ3). All groups were first asked whether participants should receive extra monetary compensation for sharing their Fitbit data in data-driven surveys (M.Q22 and M.Q23). They were then asked whether the extra monetary compensation should be an absolute (fixed) amount or a relative amount to the baseline compensation for participating in a survey (M.Q24). Then participants had to enter the additional monetary compensation amount; an absolute amount in M.Q25 and a relative amount in M.Q26. The way payments are framed can have pronounced and significant effects on people's perception of value [17, 81]. We gave participants the option to express their views both in relative and absolute amounts, to see if there is a difference and to provide more accurate budgeting and monetary compensation formulation recommendations for researchers interested in doing data-driven surveys.

Part 8 This part (M.Part 7) comprises two questions (M.Q27 and M.Q28) for evaluating the ● HYBRID group participants' perceptions regarding differences between their self-reported answers and the data collected from their Fitbit accounts (see RQ2). Fitbit data has variable accuracy based on the type of data [23, 33, 39]. Hence, understanding whether the difference is likely to be participants misremembering or not finding the relevant data in the Fitbit app UI can provide insights into how reliable data-driven surveys are for collecting self-reported data.

Part 9 In order to enable us to characterize our participant sample, this part (M.Part 8) comprises three questions (M.Q29, M.Q30, and M.Q31) for the collection of demographic information that is not available on Prolific (e.g., gender identity [90] and IUIPC [44, 66]).

B SURVEY TRANSCRIPT

To implement the logic of redirecting participants to the data-driven survey platform and back to our survey we used two separate surveys on Qualtrics.

B.1 Consent Survey Transcript

Note: [Coding rules are colored in gray (not visible to participants)] [Survey flow rules are colored in red (not visible to participants)]

[C.Part 1] [Consent survey]

[Display only if path variable missing in URL] Something has **gone wrong** with **preparing** the **survey** (Error code: 1).

Please **contact us** through **Prolific 's messaging service** if you would like to **retake the survey**.

Otherwise, click the "**Next** →" to go back to Prolific and to **return your submission without penalty**.

[Terminate]

C.Q1. What is your Prolific ID?

Please note that this response should auto-fill with the correct ID.

[(text block)]

C.Q2. **STUDY**

You are invited to participate in a study about **Fitbit smartwatch/activity tracker usage** and **survey taking experiences**. We will collect information about **how** you **use** your **Fitbit device**, such as the name of the most frequent activity you record with it and why you record that activity most often (for example, by asking why you do your most frequent activity). We will also collect information about your experiences with taking surveys.

This study is conducted and financed by Prof. Kévin Huguenin's Information Security and Privacy Lab at the University of Lausanne (UNIL), Switzerland.

PARTICIPATION CRITERIA

To be eligible for this study **you must**:

- **regularly** use a **Fitbit smartwatch/activity tracker**,
- have your Fitbit device paired with a **smartphone**,
- **use** the **official Fitbit app**,
- **regularly synchronize** your Fitbit device with the official Fitbit app,
- have a Fitbit account for **at least 6 months**,
- have **Activity data** (such as Walk, Run, Bike),
- have **Active Minutes** or **Active Zone Minutes data**, and
- have **Step data**.

OUTLINE OF THE STUDY PROCEDURE

${e://Field/consent_outline_of_the_study_procedure} [Outline shown depends on the group a participant is in.]

The survey consists of **14 to 17 questions** and should take no longer than **8-10 minutes** to complete (depending on the choices you make during the survey).

REMUNERATION

At the end of the study, you will be awarded **GBP 2.5 (~USD 3.2)** for your participation.

CONFIDENTIALITY AND DATA PROCESSING

Your answers will be recorded in a confidential and secure way. They will only be accessible to the researchers and authorized personnel from UNIL. In the case where the answers are shared with the scientific community to promote open science (open data), they will be de-identified and/or aggregated.

YOUR RIGHTS

Your participation in this study is entirely voluntary. You have the right to refuse to participate or to withdraw from the study at any time. If you withdraw from the study, your data will be deleted, and you will not be remunerated.

QUESTIONS

If you have any questions about the study, please feel free to contact the research team using the Prolific messaging service.

CONSENT

By giving your consent, you acknowledge that you are at least 18 years old. You also acknowledge that you have read the above information and that you agree to it. Please select the **Agree** option to continue. If you choose **Disagree,** you will not participate in this research survey and will not be paid.

○ Agree
○ Disagree

[✋ Terminate if response to C.Q2 is “Disagree”]

[Pre-Redirect ● MANUAL] [⤷ Display only if participant in “● MANUAL” group] ≣ **Thank you** for accepting to participate in the survey.

Please **click** "**Next** →" to be redirected to the **first part** of the **survey**.

[Pre-Redirect ● HYBRID and ● DATA-DRIVEN] [⤷ Display only if participant in “● HYBRID” or “● DATA-DRIVEN” groups] ≣ **Thank you** for accepting to participate in the survey.

Please **click** "**Next** →" to be redirected to the **first part** of the **survey**.

≣ You have been assigned to a participant group that will have to **grant access** to their **Fitbit data**.
Here you will find step by step instructions for how to do this:

1. When you click "**Next** →" you will be redirected to the data-driven survey platform (DDS). You will need to connect to your Fitbit account by clicking on **+ CONNECT**:
   [[🖼 screenshot]]
2. You will then be taken to the Fitbit sign in screen. Make sure to **sign in** with the **account** that you **use for Fitbit**:
   [[🖼 screenshot]]
3. After signing in, you will be shown the next page. It is **important** to **check Allow All at (1)**, then to click **Allow (2)**:
   [[🖼 screenshot]]
4. You will be taken back to DDS, where you can proceed to the rest of the survey by clicking on **PROCEED**:
   [🖼 screenshot]

B.2 Main Survey Transcript

Survey sections	Question numbers	Section description
—	—	Error handling
—	—	Mock survey introduction
M.Part 1	M.Q1, M.Q2, M.Q3, M.Q4, M.Q4, M.Q5, M.Q6, M.Q7	Mock survey: factual and perceptual for ● MANUAL and ● HYBRID
M.Part 2	M.Q7, M.Q8, M.Q9, M.Q10	Mock survey: perceptual for ● DATA-DRIVEN
—	—	Rest of survey introduction
M.Part 3	M.Q11, M.Q12, M.Q13	Usability (UEQ) and Checked Fitbit data
M.Part 4	M.Q14, M.Q15, M.Q16, M.Q17	Utility and Data-access issues
M.Part 5	M.Q18, M.Q19, M.Q20, M.Q21, M.Q21	Privacy and Data-access transparency
M.Part 6	M.Q22, M.Q23, M.Q24, M.Q25, M.Q26	Extra monetary compensation for sharing data
M.Part 7	M.Q27, M.Q28	Follow up mock
M.Part 8	M.Q30, M.Q31	Demographics
—	—	End

Note: [Coding rules are colored in gray (not visible to participants)] [Survey flow rules are colored in red (not visible to participants)]

[Error: No path variable] [⤷ Display only if path variable missing in URL]
≣ Something has **gone wrong** with **preparing** the **survey** (Error code: 1).

Please **contact us** through **Prolific's messaging service** if you would like to **retake the survey**.

Otherwise, click "**Next** →" to go back to Prolific and to **return your submission without penalty**.

[✋ Terminate]

[Error: Fitbit account created less than 6 months ago] [⤷ Display only if participants' Fitbit account was created less than 6 months ago]
≣ Your **Fitbit account** was **not created at least 6 months ago**.

Based on our participation criteria **you do not qualify to participate** in the survey.

If you think this is a mistake, please **contact us** through **Prolific 's messaging service**.

Click "**Next** →" to go back to Prolific and to **return your submission without penalty**.

Your **data will be deleted** and you **will not be paid**.

[✋ Terminate]

[Error: Fitbit account missing required data] [Display only if participants' Fitbit account did not have all the required data for the survey]

We could not collect the required data for this survey from your **Fitbit account**.

Based on our participation criteria **you do not qualify to participate** in the survey.

If you think this is a mistake, please **contact us** through **Prolific 's messaging service**.

Click "**Next** →" to go back to Prolific and to **return your submission without penalty**.

Your **data will be deleted** and you **will not be paid**.

[✋ Terminate]

[Mock survey introduction]

This is the **first part** of the **survey**, which consists of **questions** about **how** you **use** your **Fitbit smartwatch/activity tracker**.

[**M.Part 1**] [Mock survey: factual and perceptual for • MANUAL and • HYBRID] [Display only if participant in "• MANUAL" or "• HYBRID" groups]

⏱ Timer

[This element enables recording and managing how long a participant spends on the current page. This element is not displayed to the participant.]

For the next two questions, please think about **a typical week** over the **last six months**.

M.Q1. On a typical week, **which activity** (other than **Walk**)* is **most often** detected automatically or logged manually in your Fitbit account? [Sources: [40, 104]]

○ Sport
○ Bike
○ Run
○ Aerobic Workout
○ Swim
○ Weights
○ Workout
○ Interval Workout
○ Elliptical
○ Mountain Bike
○ Other (please specify): [(text block)]
○ Walk (select **Walk if** it is the **only logged activity**)*

M.Q2. How many **minutes** do you usually spend **exercising** in total over an entire **typical week** while **wearing** your **Fitbit**?

Note: this includes all exercise, not only the activity that is detected/logged most often.

[(text block)]

M.Q3. Approximately on **which day** over the **last six months** do you think you had your **highest step count** while wearing your Fitbit (yyyy-mm-dd)?

Note: you can use the calendar below to pick the date instead of typing one.

[📅 calendar]

[(text block)]

M.Q4. Approximately how many **steps** do you think you took on **that day** (${M.Q3})?

[(text block)]

[—Page break—]

⏱ Timer

[**M.Part 2**]

M.Q5. [Display only if participant in "● MANUAL" or "● HYBRID" groups] You reported that your most frequent activity is "**${M.Q1}**".

Explain in one or two sentences **what benefits** you get from **regularly doing** this activity.

[(text block)]

M.Q6. You reported that you do **${M.Q2} minutes** of **exercise per week** while wearing your Fitbit.

What are your **main reason(s)** or **motivation(s)** for **doing this amount** of exercise?

Select all that apply: [Sources for options: [38, 45, 51, 57]]

□ Improving health
□ Socializing
□ Improving appearance
□ Managing stress
□ Fun (I enjoy exercising)
□ Challenge
□ Training for competitions (I am an amateur athlete)
□ Training for competitions (I am a professional athlete)
□ Other (please specify): [(text block)]

M.Q7. You reported that over the past six months, your most active day was the '**${M.Q3}**', with approximately **${M.Q4}** steps taken.

Explain in one or two sentences **what was special** about **that day** that **made** your **step count higher** than other days.

[(text block)]

[Mock survey: perceptual for ● DATA-DRIVEN] [Display only if participant in "● DATA-DRIVEN" group]

⏱ Timer

M.Q8. Your Fitbit data shows that your most frequent activity is **${fitbit.most_frequent_activity}**.

Explain in one or two sentences **what benefits** you get from **regularly doing** this activity.

[(text block)]

M.Q9. Your Fitbit data shows that you do **${fitbit.average_weekly_active_time} minutes** of **exercise per week** while wearing your Fitbit.

What are your **main reason(s)** or **motivation(s)** for **doing this amount** of exercise?

Select all that apply: [Sources for options: [38, 45, 51, 57]]

□ Improving health
□ Socializing
□ Improving appearance
□ Managing stress
□ Fun (I enjoy exercising)
□ Challenge
□ Training for competitions (I am an amateur athlete)
□ Training for competitions (I am a professional athlete)
□ Other (please specify): [(text block)]

M.Q10. Your Fitbit data shows that over the past six months, your most active day was the '**${fitbit.most_active_day_last_6_months_date}**', with approximately **${fitbit.most_active_day_last_6_months_date}** steps taken.

Explain in one or two sentences **what was special** about **that day** that **made** your **step count higher** than other days.

[(text block)]

[Rest of survey introduction]

This is the **second part** of the **survey**, which consists of **questions** about your **experience** with **filling this survey**.

[**M.Part 3**] [Usability (UEQ) and Checked Fitbit data]

M.Q11. Please assess your **experience with answering** the **questions** about your Fitbit data and physical activity **during** the **first part** of the **survey**. [Source: [87, 88]]

– obstructive:supportive
– complicated:easy

- inefficient:efficient
- confusing:clear
- boring:exciting
- not interesting:interesting
- conventional:inventive
- usual:leading edge

○ 1
○ 2
○ 3
○ 4
○ 5
○ 6
○ 7

M.Q12. [⑂ Display only if participant in ● MANUAL group]
☰ Did you **check** your actual **Fitbit data** at any point **while** you were **responding** to the **previous questions**?

Please respond truthfully: *your **response** to this question **will not exclude** you from this study and it **will not affect** your **compensation**.*

○ No, I did not check my Fitbit data
○ Yes, I did check my Fitbit data

M.Q13. [⑂ Display only if participant in ● HYBRID group]
☰ Did you **check** your actual **Fitbit data** at any point (excluding the part where you gave temporary access to your Fitbit data) **while** you were **responding** to the **previous questions**?

Please respond truthfully: *your **response** to this question **will not exclude** you from this study and it **will not affect** your **compensation**.*

○ No, I did not check my Fitbit data
○ Yes, I did check my Fitbit data

[**M.Part 4**] [Utility and Data-Access issues]

M.Q14. [⑂ Display only if participant in ● MANUAL group]
☰ At the beginning of this survey, you were asked to manually respond to questions about your Fitbit data.

How **useful** would you find it if we could pre-answer the questions about your data by using your actual Fitbit data?

You would have to **grant temporary access** to your **data** like this:

[🖼 screenshot]

Your **questions looked like** this:

[🖼 screenshot]

[🖼 screenshot]

After granting temporary access to your data, your **questions would look like** this:

[🖼 screenshot]

○ Extremely useless
○ Moderately useless
○ Slightly useless
○ Neither useful nor useless
○ Slightly useful
○ Moderately useful
○ Extremely useful

M.Q15. [⑂ Display only if participant in ● DATA-DRIVEN group]
☰ At the beginning of this survey, you granted temporary access to your Fitbit data.

How **useful** did you find that some questions about your data were pre-answered by using your actual Fitbit data?

Your **questions looked like** this:

[🖼 screenshot]

Without granting temporary access, your **questions would look** like this:

[🖼 screenshot]

[🖼 screenshot]

○ Extremely useless
○ Moderately useless
○ Slightly useless
○ Neither useful nor useless
○ Slightly useful
○ Moderately useful
○ Extremely useful

M.Q16. [Display only if participant in • HYBRID group]
☰ At the beginning of this survey, you granted temporary access to your Fitbit data.

How **useful** would you find it if we had pre-answered some of the questions using your Fitbit data?

Your **questions looked like** this:

[🖼 screenshot]

[🖼 screenshot]

If we had used your data, your **questions would look** like this:

[🖼 screenshot]

○ Extremely useless
○ Moderately useless
○ Slightly useless
○ Neither useful nor useless
○ Slightly useful
○ Moderately useful
○ Extremely useful

M.Q17. [Display only if participant in • HYBRID ord • DATA-DRIVEN group]
☰ Did you **encounter** any of the following **issues** with **granting temporary access** to your **Fitbit data**?

Select all that apply:
□ I could not remember my Fitbit/Google account username and/or password
□ I had trouble understanding what I had to do
□ Other (please specify in one or two sentences): [(text block)]
□ None of the above

[M.Part 5] [Privacy and Data-access transparency]

M.Q18. [Display only if participant in • MANUAL group]
☰ **Imagine** you could **grant temporary access** to your **Fitbit data** to **personalize** this survey (for example, **automatically** pre-answering some of the questions). Consider that:

- You would need to log into your Fitbit account and grant temporary access to your Fitbit data for our survey.
- Only the needed data would be accessed.
- All data would be deleted after completing the survey, except for the data used for the personalized questions.

How **comfortable** would you be with **granting temporary access** to your **Fitbit data**?
○ Extremely uncomfortable
○ Moderately uncomfortable
○ Slightly uncomfortable
○ Neither comfortable nor uncomfortable
○ Slightly comfortable
○ Moderately comfortable
○ Extremely comfortable

M.Q19. [Display only if participant in ● DATA-DRIVEN group]
At the **beginning** of this **survey**, you **granted temporary access** to your **Fitbit data** to personalize this survey.

How **comfortable** were you with **granting temporary access** to your **Fitbit data**?
○ Extremely uncomfortable
○ Moderately uncomfortable
○ Slightly uncomfortable
○ Neither comfortable nor uncomfortable
○ Slightly comfortable
○ Moderately comfortable
○ Extremely comfortable

M.Q20. [Display only if participant in ● HYBRID group]
At the **beginning** of this **survey**, you **granted temporary access** to your **Fitbit data**.

How **comfortable** were you with **granting temporary access** to your **Fitbit data**?
○ Extremely uncomfortable
○ Moderately uncomfortable
○ Slightly uncomfortable
○ Neither comfortable nor uncomfortable
○ Slightly comfortable
○ Moderately comfortable
○ Extremely comfortable

M.Q21. [Display only if participant in ● HYBRID ord ● DATA-DRIVEN group, and they opened the privacy-transparency table on DDS]
How much did the **privacy policy** and the **summary** of the **collected data** (see the screenshot after the question for an example) **help** you **decide** to **grant temporary access** to your **Fitbit data**?
○ None at all
○ A little
○ A moderate amount
○ A lot
○ A great deal
○ I did not read them (selecting this option is fine)

Example of the collected data summary:

[screenshot]

[**M.Part 6**] [Extra monetary compensation for sharing data]

M.Q22. [Display only if participant in ● MANUAL group]
Imagine you could have **granted** direct temporary **access** to your **Fitbit data** (as explained previously).

Do you **think** you **should receive extra compensation** for **granting temporary access** to your **Fitbit data**?

Note: *your* ***response*** *to this question* ***will not influence*** *the* ***payment you receive*** *for this study.*
○ No
○ Yes

M.Q23. [Display only if participant in ● HYBRID or ● DATA-DRIVEN group]
Do you **think** you **should receive extra compensation** for **granting temporary access** to your **Fitbit data**?

Note: *your* ***response*** *to this question* ***will not influence*** *the* ***payment you receive*** *for this study.*
○ No
○ Yes

M.Q24. [Display only if response to M.Q23 is "Yes"]
Do you **think** the **extra compensation** should be a **relative** amount (that is, some percentage more) or an **absolute** amount (that is, a fixed amount more)?
○ Relative amount (that is, some percentage more)
○ Absolute amount (that is, a fixed amount more)

M.Q25. [Display only if response to M.Q24 is “Absolute amount (that is, a fixed amount more)”]
For this study, you would **normally** be **paid 1 GBP**.

How much more would you like to get?

Here is how much you would get in such a scenario:
1 GBP + 0 GBP = **1 GBP** [Calculation updates based on participant’s input]

[(text block)]

M.Q26. [Display only if response to M.Q24 is “Relative amount (that is, some percentage more)”]
For this study, you would normally be paid **1 GBP.**

How much of a **percentage increase** should you get?

Here is how much you would get in such a scenario:
1 GBP + 0% = **1 GBP** [Calculation updates based on participant’s input]

[(text block)]

[**M.Part 7**] [Follow up mock] [Display only if participant in • HYBRID group]

M.Q27. [Display only if response to M.Q1 is not equal to the most frequent activity in the participant’s Fitbit account]
You reported that your most frequent activity is "**${self_reported_most_frequent_activity}**".

Your Fitbit data shows that your most frequent activity is ${most_frequent_activity}.

Please explain in one or two sentences **why** you think there **is a difference** between what **you remember** and what **your Fitbit data shows**.

*Note: this **does not exclude** you from this study and it **does not affect** your **compensation**.*

[(text block)]

M.Q28. [Display only if response to M.Q3 is not equal to the date of the day with the highest step count over the last 6 months in the participant’s Fitbit account]
You reported that over the past six months, your most active day was the '**${self_reported_most_active_day_date}**', with approximately **${self_reported_most_active_day_steps}** steps taken.

Your Fitbit data shows that over the past six months, your most active day was the '**${fitbit.hghst_stps_lst_6_mnths_date}**', with approximately **${fitbit.hghst_stps_lst_6_mnths_stps}** steps taken.

Please explain in one or two sentences **why** you think there **is a difference** between what **you remember** and what **your Fitbit data shows**.

*Note: this **does not exclude** you from this study and it **does not affect** your **compensation**.*

[(text block)]

[**M.Part 8**] [Demographics]

M.Q29. What is your gender? [Source: [90]]
□ Woman
□ Man
□ Non-binary
□ Prefer to self-describe [(text block)]
□ Prefer not to disclose

M.Q30. How many **surveys** do you **fill** on average per year **through** the **Prolific** platform/service?

[(text block)]

M.Q31. Please indicate to what extent you **agree** with each of the following statements. [Source: IUIPC-8 [44, 66]]

- Consumer online privacy is really a matter of consumers’ right to exercise control and autonomy over decisions about how their information is collected, used, and shared
- Consumer control of personal information lies at the heart of consumer privacy
- Companies seeking information online should disclose the way the data are collected, processed, and used
- A good consumer online privacy policy should have a clear and conspicuous disclosure
- It usually bothers me when online companies ask me for personal information
- When online companies ask me for personal information, I sometimes think twice before providing it
- It bothers me to give personal information to so many online companies
- I am concerned that online companies are collecting too much personal information about me

○ Strongly disagree
○ Disagree
○ Moderately disagree
○ Neither agree nor disagree
○ Moderately agree
○ Agree
○ Strongly agree

[End]

Thank you for **participating** in this survey!

Please **click** "**Next** →" to be **redirected** back to **Prolific** to be **eligible** to be **paid**.

[Display only if participant in ● HYBRID or ● DATA-DRIVEN group]

DDS revokes access automatically once you are redirected to this survey.

To verify that DDS' access to your data has been revoked, you can do the following:

1. Go to: https://www.fitbit.com/settings/applications
2. Make sure that 'Usage Study by University of Lausanne - ISP Lab' **is not** in the list of **Applications**.

If 'Usage Study by University of Lausanne - ISP Lab' is in the list, please do the following to revoke access:

1. Click on '**Revoke Access**' next to 'Usage Study by University of Lausanne - ISP Lab' in the list of **Applications**:

   [🖼 screenshot]

2. Click 'Confirm':

   [🖼 screenshot]

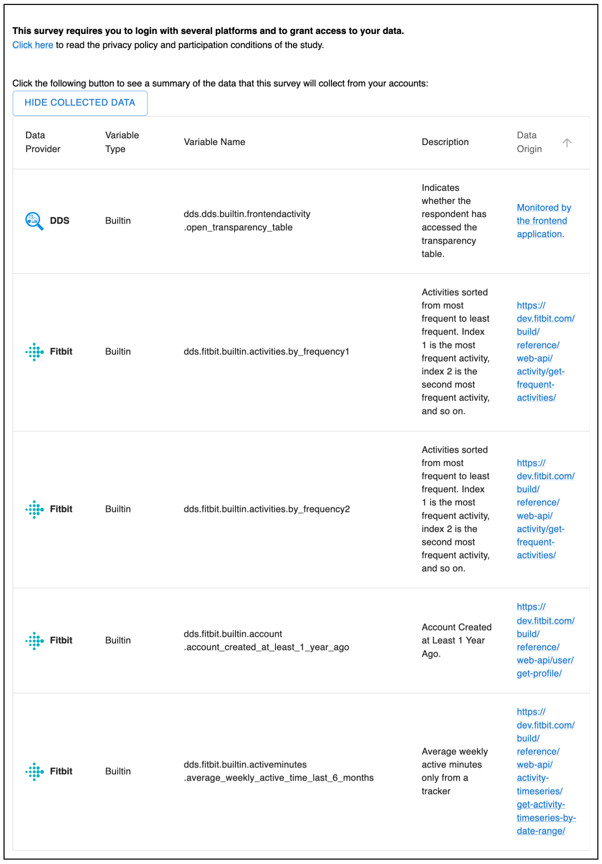
C PARTICIPANTS' DEMOGRAPHICS TABLE

Table 5. Participants' demographics.

	Statistics	● MANUAL	● DATA-DRIVEN	● HYBRID	Overall
Age	*M* ± *SD*	44.1 ± 11.4	43.4 ± 13.1	42.8 ± 12.6	43.5 ± 12.4
	Min — Max	20 − 70	19 − 72	20 − 76	19 − 76
Gender	Woman	47.0%	69.0%	72.0%	62.7%
	Man	52.0%	30.0%	27.0%	36.3%
	Non-Binary	0.0%	1.0%	1.0%	0.7%
	Did Not Answer	1.0%	0.0%	0.0%	0.3%
IUIPC-8 (btw. 1 and 7)	*M* ± *SD*	5.8 ± 0.8	5.7 ± 0.9	5.8 ± 0.8	5.7 ± 0.8
Ethnicity	Asian	9.0%	11.0%	7.0%	9.0%
	Black	13.0%	6.0%	10.0%	9.7%
	Mixed	4.0%	7.0%	5.0%	5.3%
	Other	1.0%	1.0%	3.0%	1.7%
	White	73.0%	75.0%	75.0%	74.3%
Physical Activity	Most Freq.	Run (30%)[SR]	Sport (18.8%)[F]	Walk (54%)[SR,†] Run (16.7%)[F,†]	
Weekly Exercise Time	*M* ± *SD*	224.8±228.8[SR]	266.2 ± 337.0[F]	136.0 ± 134.8[SR,†] 278.9 ± 416.9[F,†]	
Most Act. Day Step Count	*M*±*SD*	16010.7 ± 11824.0[SR]	17587.4 ± 9235.3[F]	12398.7 ± 6408.9[SR,†] 16130.0 ± 7601.5[F,†]	
Motivation to Exercise	Most Freq.	Improving health (91%)	Improving health (88%)	Improving health (85%)	
	2nd Most Freq.	Managing stress (60%)	Improving appearance (56%)	Managing stress (58%)	

[SR] Calculated using self-reported values from the mock-survey (M.Q1, M.Q2, M.Q4, M.Q6).
[F] Calculated using the participants' Fitbit data.
[†] The discrepancies between these values are analyzed in Section 4.2.

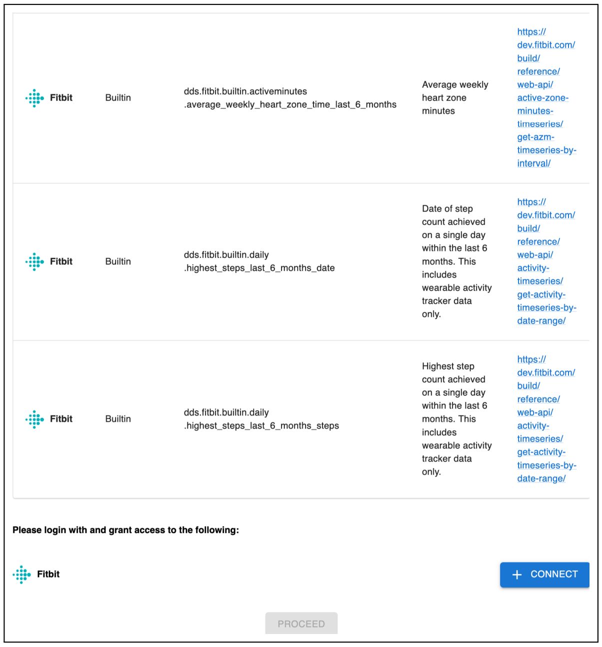
D PRIVACY TRANSPARENCY TABLE FROM DDS



This survey requires you to login with several platforms and to grant access to your data.
Click here to read the privacy policy and participation conditions of the study.

Click the following button to see a summary of the data that this survey will collect from your accounts:

Data Provider	Variable Type	Variable Name	Description	Data Origin
DDS	Builtin	dds.dds.builtin.frontendactivity .open_transparency_table	Indicates whether the respondent has accessed the transparency table.	Monitored by the frontend application.
Fitbit	Builtin	dds.fitbit.builtin.activities.by_frequency1	Activities sorted from most frequent to least frequent. Index 1 is the most frequent activity, index 2 is the second most frequent activity, and so on.	https://dev.fitbit.com/build/reference/web-api/activity/get-frequent-activities/
Fitbit	Builtin	dds.fitbit.builtin.activities.by_frequency2	Activities sorted from most frequent to least frequent. Index 1 is the most frequent activity, index 2 is the second most frequent activity, and so on.	https://dev.fitbit.com/build/reference/web-api/activity/get-frequent-activities/
Fitbit	Builtin	dds.fitbit.builtin.account .account_created_at_least_1_year_ago	Account Created at Least 1 Year Ago.	https://dev.fitbit.com/build/reference/web-api/user/get-profile/
Fitbit	Builtin	dds.fitbit.builtin.activeminutes .average_weekly_active_time_last_6_months	Average weekly active minutes only from a tracker	https://dev.fitbit.com/build/reference/web-api/activity-timeseries/get-activity-timeseries-by-date-range/



Data Provider	Variable Type	Variable Name	Description	Data Origin
Fitbit	Builtin	dds.fitbit.builtin.activeminutes .average_weekly_heart_zone_time_last_6_months	Average weekly heart zone minutes	https://dev.fitbit.com/build/reference/web-api/active-zone-minutes-timeseries/get-azm-timeseries-by-interval/
Fitbit	Builtin	dds.fitbit.builtin.daily .highest_steps_last_6_months_date	Date of step count achieved on a single day within the last 6 months. This includes wearable activity tracker data only.	https://dev.fitbit.com/build/reference/web-api/activity-timeseries/get-activity-timeseries-by-date-range/
Fitbit	Builtin	dds.fitbit.builtin.daily .highest_steps_last_6_months_steps	Highest step count achieved on a single day within the last 6 months. This includes wearable activity tracker data only.	https://dev.fitbit.com/build/reference/web-api/activity-timeseries/get-activity-timeseries-by-date-range/

Please login with and grant access to the following:

Fitbit

Fig. 5. Privacy transparency table shown to participants before they grant access to their data on DDS.

E CODEBOOK FOR M.Q27 and M.Q28

Table 6. Codebook for M.Q27 and M.Q28.

Code	Frequency	Example Quote
Discrepancy in Activity Type (M.Q27)		
Fitbit could not detect the activity type accurately	$n = 48$	*"I would guess it is a misread, as I have not used an elliptical in the past 6 months."*
Users call the activity differently than how Fitbit calls it	$n = 7$	*"I'm not very fluent in fitness vocabulary. I don't always know what to label my workouts."*
Participants admitted their oversight	$n = 6$	*"I didn't realize."*
Forgetfulness/memory issue	$n = 5$	*"I have a horrible memory."*
Participants misunderstood the question	$n = 3$	*"I thought it said to choose one other than walking."*
Misunderstanding of Fitbit tracking capabilities	$n = 1$	*"I couldn't think of another activity that Fitbit usually tracks for me other than horseback riding, but I haven't been riding lately."*
Others, including irrelevant answers	$n = 4$	—
Discrepancy in Active Date (M.Q28)		
Forgetfulness/memory issue	$n = 47$	*"I am busy a lot of days, but can't really remember day to day how many steps I take."*
Participants admitted their oversight	$n = 14$	*"I looked back in the app, but I was going quickly and must have missed that day. Looking at the calendar, that was a book shelving day!"*
Fitbit could not measure accurately	$n = 3$	*"I don't normally walk very much in July because it's too hot. I went on one walk that day, 2.2 miles. I cut grass that day, and the Fitbit counted movement it shouldn't have. I walked much more in September and October."*
Participants misunderstood the question	$n = 3$	*"I think I misread the question and thought you asked about the past month only."*
Because of Fitbit's UI limitations	$n = 1$	*"The Fitbit data I was looking at was on my Fitbit app on my smartphone in graph form, not showing the actual number; I chose the graph that appeared the highest and then zeroed in on that day to again just get an approximate number. I don't pay for the subscription, so my data access may be limited."*
Others, including irrelevant answers	$n = 9$	—