




Designing Reflective Derived Metrics for Fitness Trackers

MARIT BENTVELZEN , Utrecht University, the Netherlands

JASMIN NIESS , University of St. Gallen, Switzerland

PAWEŁ W. WOŹNIAK , Chalmers University of Technology, Sweden

Personal tracking devices are equipped with more and more sensors and offer an ever-increasing level of accuracy. Yet, this comes at the cost of increased complexity. To deal with that problem, fitness trackers use *derived metrics*—scores calculated based on sensor data, e.g. a stress score. This means that part of the agency in interpreting health data is transferred from the user to the tracker. In this paper, we investigate the consequences of that transition and study how derived metrics can be designed to offer an optimal personal informatics experience. We conducted an online survey and a series of interviews which examined a health score (a hypothetical derived metric) at three levels of abstraction. We found that the medium abstraction level led to the highest level of reflection. Further, we determined that presenting the metric without contextual information led to decreased transparency and meaning. Our work contributes guidelines for designing effective derived metrics.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**.

Additional Key Words and Phrases: personal informatics, metrics, derived metrics, reflection, fitness trackers




ACM Reference Format:

Marit Bentvelzen , Jasmin Niess , and Paweł W. Woźniak . 2022. Designing Reflective Derived Metrics for Fitness Trackers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 158 (December 2022), 19 pages. <https://doi.org/10.1145/3569475>

1 INTRODUCTION

‘Your stress level is 77%.’ This is an example message which one might receive from a modern fitness tracker. The message represents a new trend in personal informatics devices—trackers have transitioned from providing overviews of measured, objective data (e.g. average heart rate or step counts) to *interpreted data* (e.g. stress or energy levels). This change is apparent both in the range of new metrics available in new fitness tracker models [11] and the marketing narrative around fitness tracking. The new (at time of writing) Fitbit Charge 5 features six ‘reasons you’ll love Charge 5’¹ and four of these reasons are interpreted data: Daily Readiness, Stress Management, Heart Health and Health Metrics. One reason for the need to use interpreted data is the fact that leading fitness trackers on the consumer market already include advanced sensors, which generate an abundance of data points. Thus, processing tracker data is inevitable to avoid complexity [22, 26]. Yet, processing physiological data before showing it to the user inherently involves aggregating and processing sensor data using computational tools and delegating part of the data interpretation to those who design the processing algorithms. Consequently, it is a challenge for Human-Computer Interaction (HCI) to understand how such *derived metrics*

¹<https://www.fitbit.com/global/se/products/trackers/charge5>

Authors’ addresses: Marit Bentvelzen , Utrecht University, Utrecht, the Netherlands, m.bentvelzen@uu.nl; Jasmin Niess , University of St. Gallen, St. Gallen, Switzerland, jasmin.niess@unisg.ch; Paweł W. Woźniak , Chalmers University of Technology, Gothenburg, Sweden, pawelw@chalmers.se.



This work is licensed under a Creative Commons Attribution International 4.0 License.

© 2022 Copyright held by the owner/author(s).

2474-9567/2022/12-ART158

<https://doi.org/10.1145/3569475>

(we propose adopting this term and use it in this paper²) can be designed in ways which effectively benefit the users' wellbeing.

Simultaneously, the HCI field has studied the personal informatics experiences created by fitness trackers and recognises the role of metrics as a key design dimension of such experiences. Epstein et al. [18] concluded that selecting tools (and, consequently, metrics) was a key step in starting a personal informatics experience. Niess and Woźniak [48] showed how users translated their qualitative goals into quantitative metrics to achieve a meaningful tracking experience. These investigations and others show that effective metrics are core to a positive tracking experience. Further, not only what the metrics represent, but also how the data is presented to the user affects the tracking experience. Past work has shown that visual representations of tracker data can affect user experience [32] and how the users reflect upon the data [46]. Thus, exploring how to present derived metrics and the data behind them in ways which foster optimal user benefit is key to improving personal informatics experiences. As the complexity and computing power of personal trackers increases, we need to find ways to optimally communicate the sensor information to users and ensure that they can have more agency by making more informed decisions about their wellbeing.

To understand user perceptions of derived metrics and learn how such metrics can support wellbeing, this paper contributes a two-step study of different forms of derived metrics. We conducted an online survey and an interview study which examined three ways of presenting a derived metric for health at three levels of abstraction (i.e. metrics presented with varying levels of detail and vagueness, representing different levels of construal [54]). We determined that a medium level of abstraction produced the most reported reflection in users. Based on the data gathered, we conclude that this was due to the fact that a medium level of abstraction allowed for construal alignment, i.e. understanding the metric at a level of abstraction which enables reflection and identifying underlying reasons for the metric. This implies that a well-designed derived metric provides enough contextual information to avoid confusion and, simultaneously, serves to simplify the data gathered about one's wellbeing.

This paper contributes the following: (1) an online survey comparing three levels of abstraction of a health metric for fitness trackers; (2) an interview study of user attitudes toward derived health metrics at different levels of abstraction and (3) guidelines for designing and presenting derived metrics in fitness trackers.

2 RELATED WORK

In this section, we review past work on reflection in personal informatics systems, then discuss related work on metrics in personal informatics as well as derived metrics used by commercial fitness trackers. We then discuss relevant previous research focusing on abstraction in presenting information.

2.1 Reflection in Personal Informatics

Personal informatics systems enable users to come to an understanding of their health and wellbeing through automatically collected personal data. By reflecting on personal data, a user can come to an understanding of patterns and trends, which can lead to more knowledge about oneself [17]. Several models have been proposed that describe the user's journey in using personal informatics systems. Epstein et al. [18] proposed the *Lived Informatics Model of Personal Informatics*, which is an extension of Li's *Stage-Based Model of Personal Informatics Systems* [36]. The Lived Informatics Model consists of four stages: deciding to track, selecting tools, tracking and acting (which is considered to be an ongoing process of collection, integration, and reflection), and lapsing. This model was later extended by Bentvelzen et al. [3] who further dissected the reflection stage in the personal informatics process. Their *Technology-Mediated Reflection Model (TMRM)* describes user behaviors and practices in the reflection phase of a personal informatics journey. The model shows how users enter, exit and stay in the reflection phase. In the TMRM, reflection is considered to be a temporary state—a dynamic process in which a

²We call metrics which represent physiological sensor data 'measurements' for contrast.

user constantly adapts their tracking experience to their evolving needs. The temporal and conceptual cycles are essential in this process. These cycles show how users' needs and perspectives evolve throughout their engagement with their tracker. While the TMRM offers an understanding of reflection on a meta-level, it remains an open question how, on a more detailed level, the metrics used to communicate data to a user affect facilitated reflection. Our work aims to gain a deeper understanding of how users perceive health and wellbeing metrics of personal informatics systems. Consequently, we aim to build insights about how such metrics should be designed to foster reflection.

2.2 Metrics in Personal Informatics

Early examples of personal informatics studies often focused on analysing or designing for a single metric. For instance, a multitude of personal informatics systems used steps as the studied metric [17]. Examples of such systems include *UbiFit Garden* [7] and *Fish'n'Steps* [39], which both aimed to encourage physical activity. Other personal informatics studies focused on heart rate [30, 40, 45], weight [13, 51] and calories [55].

Previous work indicates that the more complex or abstract the metrics become, the more users struggle to make sense of it. To illustrate, a wide variety of previous personal informatics studies focused on sleep (e.g. [34]). These studies range from methods for automatically tracking sleep [35] through accuracy of sleep tracking by commercial fitness trackers [38], guided self-experiments [10] to the credibility of sleep data [37]. Liang and Ploderer [37] demonstrated that participants found it difficult to assess the credibility of sleep tracking devices. Their study identified two *sleep fallacies* that complicated a user's assessment of the credibility of the device: (1) sleep fallacy and (2) black-box fallacy. Because sleep is a complex phenomenon, participants often misinterpreted the different sleep-stages. Furthermore, the sleep tracking devices often remain a black box, which makes it impossible for users to fully understand how the metrics are measured and processed by the device. These fallacies in turn complicate the assessment of a device's credibility. The sleep example shows how a personal tracking metric can carry inherent complexity. Dealing with and understanding that complexity is the topic of our work.

Similarly, such complexities were also found in studies that focused on stress. A study by Ding et al. [11] inquired how users of fitness trackers use the automatically tracked stress data in practice. Their study demonstrated that participants struggled to meaningfully engage with this data; some only had a limited understanding of their stress data, while others were confused. The authors mention that participants experienced a mismatch in the physiological stress measured by the device and their personal conceptualisations of stress. This, in turn, led to difficulties for users to make sense of their stress data and use it in a meaningful way [11].

The examples above illustrate that some fitness tracker metrics are more complex than others. In other words, some of the metrics which fitness trackers use can be directly experienced and quantified with counting (e.g. steps and heart rate), while other metrics are more abstract, meaning that they can be interpreted in multiple ways (e.g. personal stress score) and cannot be measured directly (e.g. sleep). Such metrics, therefore, need to be *derived* from other metrics (e.g. movement and heart rate). To date, metrics have primarily been studied in isolation, focusing on understanding a single metric at a time while placing less emphasis on how their level of abstraction influences users' perceptions of them. Thus, there is a need to interpret metrics not only as data sources which need to be presented optimally, but also as complex artefacts with internal structures which can be designed.

2.3 Derived Metrics in Fitness Trackers

Over the years commercial fitness trackers have introduced new metrics that inform users about their health and wellbeing. Early models were often limited to easily measurable metrics such as the number of steps, heart rate and sleep quantities. Recently, fitness trackers have begun to incorporate metrics such as *Stress Levels* [28], *Stress Management Scores* [23], *Sleep Scores* [24, 56], and a personalised *Recovery Score* [56]. Garmin implemented

a *Stress Levels* metric into fitness trackers in 2017, followed by the introduction of the *Body Battery* [27] metric a year later. The Body Battery informs users about their changing energy levels during the day, using a scale of 0–100 (0 denotes a low energy level and 100 denotes a high energy level). Garmin does not share details on the calculation of the metric, but the Garmin Connect application informs users that their Body Battery score is derived from heart rate signals such as Heart Rate Variability (HRV), stress score, sleep quality, and activity. Garmin’s Stress Level is derived from HRV and it shows users a value of 0–100, where the score is mapped to four categories: rest (0–25), low (26–50), medium (51–75) and high (76–100). Fitbit on the other hand uses a *Stress Management Score*, which is also a score of 0–100, yet a higher number in this case shows users that they are properly managing their stress. Similar to Garmin, Fitbit only vaguely describes the calculation of this metric. The company shares that it is based on three metrics: *responsiveness*, *exertion balance* and *sleep patterns*. The Garmin application and Fitbit are examples of the proliferation and general lack of transparency of derived metrics in modern fitness trackers. While tracker manufacturers provide a limited explanation of how the metrics are obtained, it also remains unknown how users interpret such metrics. Further, the fact that users rarely review direct sensor data but rather rely on processed feedback implies that a part of the data interpretation process is delegated to a metric processing algorithm designed by the tracker manufacturer. As consequence, there is a need to understand if and how such outsourcing can be achieved in a way that benefits the users’ wellbeing.

To summarise, the derived metrics introduced by commercial fitness trackers are often not clearly defined, leaving room for the user to interpret and conceptualise it. This ambiguity regarding what a derived metric (e.g. Body Battery) means and how it is measured makes the metric more abstract than intermediate metrics such as steps or heart rate. It remains unclear how this difference in abstraction influences a user’s understanding of their health and wellbeing. In this work, we endeavor to investigate how the different levels of metric abstraction are perceived and how this potentially impacts upon reflection, transparency and meaning.

2.4 Abstraction in Presenting Information

Previous work by Bentvelzen et al. [3] demonstrated that users of fitness trackers frequently changed the level of abstraction that they interacted with on their tracking data. This enables effective reflection as it introduces a change of perspective [25]. Conceptually, this behavior is explainable by the Construal Level Theory (CLT) by Trope et al. [54]. This theory differentiates between high level, or relatively abstract, and low level, or relatively specific, construal; the level being identified as relevant for further cognitive processing of information. The theory notes that physical activities can be construed either in a more abstract (e.g. focusing on the bigger picture) or more concrete manner (e.g. focusing on how an activity is performed). Hence, the CLT could offer a lens through which we can view the abstraction levels of metrics. Further, by varying the level of abstraction (construal level) of different metrics, fitness trackers allow users to conceptually align the information the tracker offers them to their needs, resulting in facilitated reflection [3]. However, it remains unclear how users of fitness trackers perceive the derived metrics and how the level of abstraction influences their reflections or understanding of their data.

3 METHOD

We investigated derived metrics using a mixed-method approach. First, we conducted an online survey to quantitatively compare user perceptions of derived metrics presented at different levels of abstraction. This allowed us to gain an understanding of preferences and key differences in designing derived metrics. Having observed significant differences between the metric presentations, we then decided to further study the reasons behind the differences. To this end, we conducted a series of interviews where users reflected on different designs for derived metrics.

We used the same example derived metric in both studies—overall health. We decided upon a health metric because it represented a complex, dynamic concept with many possible interpretations and sources of measurement [4]. Thus, exploring health allowed us to investigate the possible ambiguity and loss of information associated with using derived metrics. Further, an overall health metric is not currently present in major commercial fitness trackers, which allowed us to avoid familiarity bias, whereas metrics such as stress, energy or body battery could have been familiar to some users. Understanding and pursuing health is an important motivation for engaging in personal informatics [18], making it a good choice for participants to effectively contextualise the data and relate it to their experiences. Thus, the example used in our research represents the same class as current commercial metrics which simplify a complex concept, e.g. stress, to a simple score. This allows us to explore probable, yet fictional usage scenarios of fitness trackers, in the spirit of design fictions [1]. Just like real-life derived metrics, our health metric is arbitrarily derived by designer decision. Just like derived metrics in current fitness trackers, the fictional health metric is by no means correct [41]. Thus, while we use a health score as a prime example of a derived metric, we do not advise designers and or researchers to deploy such a score.

The prototypes were designed to resemble current mobile applications for fitness trackers in order to avoid novelty bias and facilitate comprehension based on past experiences. To further relate the prototypes to existing fitness tracker experiences, we used existing derived metrics as contributing to the health metric. We decided not to use the participants' data for the experiment, but opted for a vignette study where they were asked to imagine that a given data set belonged to them. This was motivated by two aspects. First, past research has shown that differences in how fitness data is presented can lead to possible negative experiences and stimulate negative thought cycles [57]. This would constitute an ethical pitfall for our study. We thus avoided this issue while still presenting a positive but not perfect health score to all participants. Second, presenting different health scores to different participants would mean that the primary factor contributing to their perception of the prototype would be their subjective assessment of the score and the resulting affective response. Showing the same data to all participants allowed us to focus on the presentation and understanding of the metric and not a particular score.

The three prototypes used in the study are presented in Figure 1. In order to study how users understand derived metrics and relate them to their lives [48], we built prototypes on three levels of abstraction. We expected that users would conceptualise derived metrics in ways similar to tracker measurements studied in past research. This implies that such metrics would need to be transparent [19], relatable to qualitative observations [48] and foster reflection [3]. In line with research by Bentvelzen et al. [3], an effective metric should allow users to interact with it at different construal levels, facilitating construal alignment. We operationalise this requirement through designing the health metric at three levels of abstraction: low, medium and high. The levels of abstraction differ in terms of the amount of information and explanation about the health metric. The low abstraction version provides extensive information about measurements and metrics contributing to the health score along with a textual information about how heart rate data is analysed for the purpose. The medium abstraction level contains a subset of the additional metrics while the high abstraction version is only the health score. Consequently, the three speculative designs in our study represent three ways in which a derived metric can be presented to users with varying complexity, level of information and level of abstraction. The particular designs were constrained by the amount of information which fits on a smartphone screen.

The data collected in both studies was anonymised and no identifying information was collected. As such, both studies were exempt from an ethics approval process as per the rules in place at the conducting institution at the time of the study.

4 STUDY I: ONLINE SURVEY

We first explored derived metrics in a qualitative between-subject study where we compared the three versions of presenting the health metric.

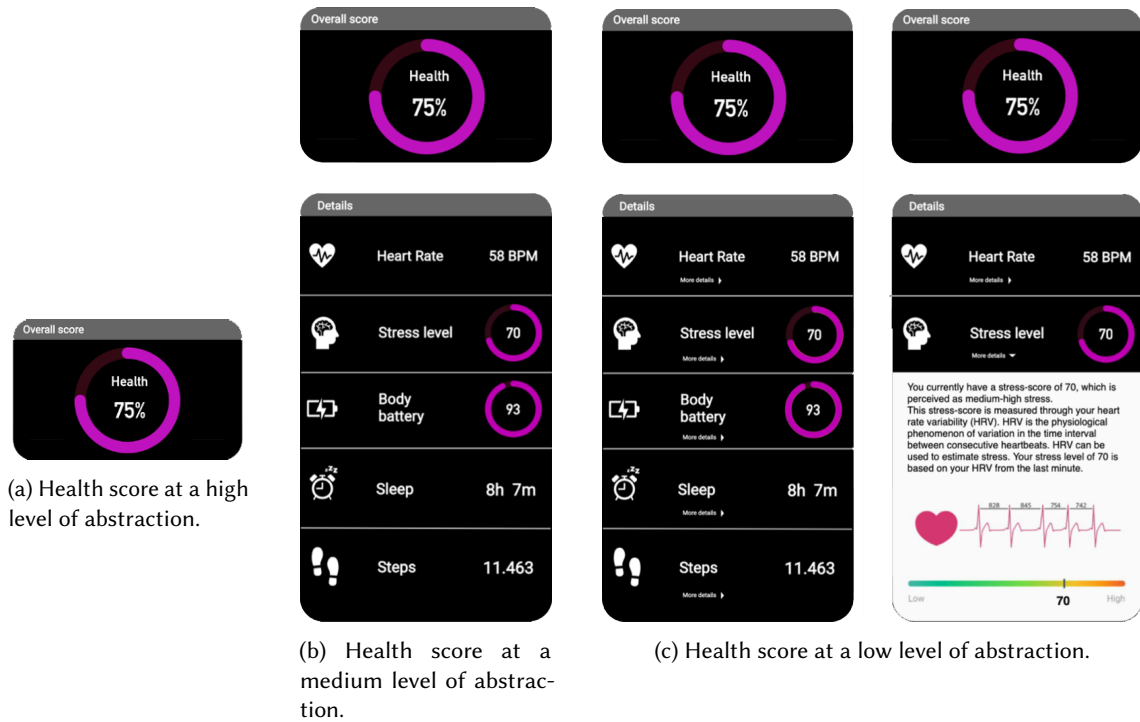


Fig. 1. Three app screen prototypes, at three different levels of abstraction, used in both studies. The design of the visualisations was inspired by current fitness tracker designs.

4.1 Participants

We recruited $n = 228$ participants, aged $M = 36.40$, $SD = 11.10$, via Amazon Mechanical Turk. The participants were mainly located in the European Union, Canada or the USA and consisted of 132 males, 95 females and 1 non-binary participant. We required a minimum of 1000 completed tasks with an acceptance rate of at least 95% to be eligible for the survey. The survey took $M = 5.15min$, $SD = 2.23min$ to complete. This method replicated past studies in personal informatics [18, 48]. The participants received USD 1 for completing the survey, which corresponds to the hourly rate for experiment participation prescribed by the conducting institution. Most of the participants reported having fitness tracker experience. Seventy-four participants had been actively tracking activities for more than a year and 19 never used a fitness tracker. Full demographic data is available in the supplementary material.

4.2 Survey Content

We used the Qualtrics XM platform to implement the survey. The survey started with gathering basic demographic data. Next, we assessed the participants' propensity to reflect on their data, by administering the reflection part of the Rumination-Reflection Questionnaire (RRQ) by Trapnell and Campbell [53]. This allowed us to later control for the participants' *Trait Reflection*, by asking participants to rate their agreement to statements such as 'I love analysing why I do things.'. Afterwards, participants were randomly assigned to one of the conditions. This was followed by an introductory video, created to increase immersion in the vignette. The video depicted a man walking in the forest and then reviewing his health data in an app after the walk. This allowed us to fully utilise

the vignette format for the study—clearly defining a context in which the participant’s views are requested, fostering a rational assessment [21]. The full survey and video are available in the supplementary material. After the video, the survey showed the application screen prototype at the assigned level of abstraction. We then proceeded with administering all the measures, listed below, constantly displaying the application prototype on top the questions.

4.3 Conditions and Measures

The study was a between-subjects design with three conditions, representing the three levels of abstraction of the prototype: LOW, MEDIUM and HIGH. In each condition, we measured the following.

4.3.1 Reflection. We used the Technology-Supported Reflection Inventory (TSRI [2]) to measure how much reflection the different prototypes produced in participants. The TSRI offers overall reflection scores and three subscales which may identify the sources of reflection: insight, exploration and comparison. As suggested by Bentvelzen et al. [2], we measured trait reflection before using the TSRI to control for character differences among participants. The TSRI asks the participants to rate their agreement with a set of nine statements which concern different aspects of technology-supported reflection which may be prompted by personal data: insight (e.g. *Using the system has led to a wake-up call to make changes in my life.*), exploration (e.g. *I enjoy exploring my data with the system.*), and comparison (e.g. *I reflect on my data in the system with others.*).

4.3.2 Transparency. We also wanted to assess if the users understood the data presented in the prototypes and whether they perceived them as transparent. There is a lack of measurement instruments in this area and the focus is primarily on trust, e.g. [42]. Thus, we opted for a not fully validated metric for transparency developed by Cramer et al. [8], which was also previously used in one personal informatics paper [57]. This instrument is a one-dimensional, rapid way of assessing perceived transparency. We rephrased the two statements from the original paper [8] to include the content of the studied prototype: *I understand what the health app bases the health information on;* *I understand how the health app calculated my health information.*

4.3.3 Meaning. Past work has indicated that creating meaning should be a key design goal for personal informatics [15, 48, 50] and even interactions with technology at large [43]. Consequently, we wanted to quantify how the different conditions possibly produce an experience of meaning. To that end, we applied the brief version of a scale developed by Huta and Ryan [33] to assess meaning. This scale was previously used to measure meaning across different categories of experiences. We chose this instrument specifically because the theories underlying the scale focus on self-improvement, which is also associated with personal informatics experiences. The scale asks participants to rank their agreement with four statements, prompting them to assess whether the health score (in the case of our experiment) is *meaningful*, *valuable*, *precious* and *full of significance*.

4.4 Results

We used analysis of variance (ANOVA) procedures to analyse the data collected in the survey. All reported p-values were Bonferroni-Holm corrected.

4.4.1 Reflection. The grand mean of TSRI scores was $M = 35.00$, $SD = 5.01$. As shown in Figure 2a, the highest TSRI scores were recorded for the MEDIUM ($M = 36.08$, $SD = 4.27$) condition, followed by LOW ($M = 34.95$, $SD = 5.10$) and HIGH ($M = 34.01$, $SD = 5.40$). We conducted a one-way ANCOVA to determine the effect of the level of abstraction of the application screen on the TSRI score, controlling for trait reflection. There was a significant effect $F_{2,224} = 3.045$, $p < .05$, $\eta^2 = .05$. The covariate was significant, $p < .001$. Post-hoc tests using Tukey HSD showed that there were significant differences between the pair MEDIUM–HIGH at $p < .01$.

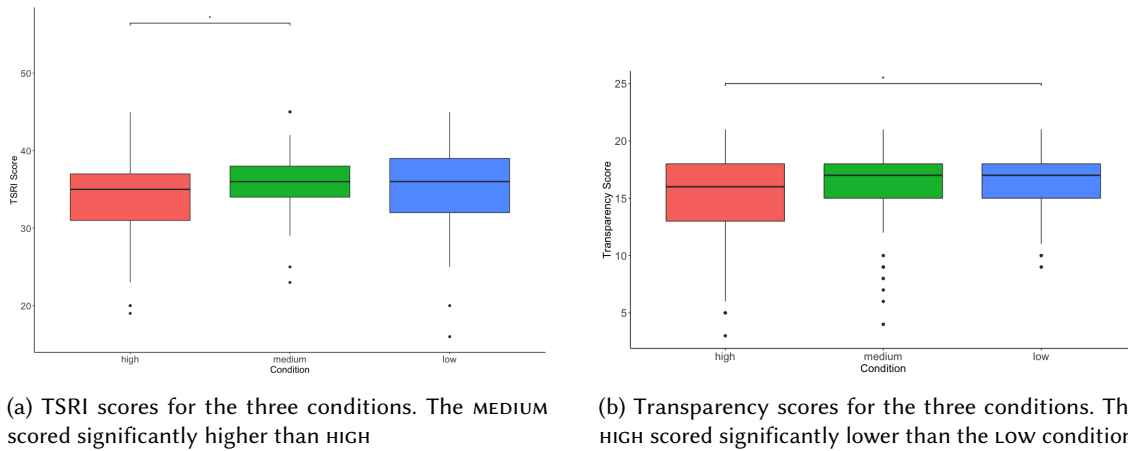


Fig. 2. Box plots for reported reflection potential and transparency across the three conditions.

Additionally, we conducted the analysis for all the subscales of the TSRI. The effects were the same as for the overall score. The full data set is available in the auxiliary material, along with the analysis.

4.4.2 Transparency. Next, we analysed the transparency scores obtained in the survey. The data analysis showed that the grand mean of transparency scores was $M = 15.61$, $SD = 3.81$. The LOW abstraction version of the prototype was reported to have the highest mean transparency, $M = 16.35$, $SD = 2.81$, followed by MEDIUM ($M = 15.62$, $SD = 3.89$) and HIGH ($M = 14.84$, $SD = 4.44$). Figure 2b shows the result. A one-way ANOVA showed a significant effect of the condition on transparency scores, $F_{2,225} = 3.07$, $p < .05$, $\eta^2 = .03$. Using Tukey HSD, we determined that only the pair LOW–HIGH was significantly different, at $p < .05$.

4.4.3 Meaning. We applied a similar procedure to analyse the effect of the conditions on perceived meaning of the data. The grand mean of meaning scores was $M = 21.87$, $SD = 4.04$. Participants exposed to the MEDIUM condition reported the highest perception of meaning, $M = 22.36$, $SD = 3.71$ with LOW ranked second, $M = 22.26$, $SD = 3.68$ and HIGH receiving the lowest scores, $M = 20.32$, $SD = 4.56$. The results can be seen in Figure 3. We conducted a one-way ANOVA and obtained a significant effect of the condition, $F_{2,225} = 6.28$, $p < .01$, $\eta^2 = .05$. In post-hoc tests using Tukey HSD, we found that the pairs LOW–HIGH and MEDIUM–HIGH were significantly different, both at the $p < .01$ level.

4.5 Interpreting the Quantitative Results

We observed significant differences between the conditions across the three measures. In all three cases, the effect size is small. The results show that the level of abstraction does affect the users' perception of presented wellbeing data across the three dimensions examined. There is also partial evidence that the medium abstraction level may be most beneficial to users. We note that the small effect sizes may be due to our choice to use a vignette design—as the data belonged to a hypothetical person, it elicited less extreme emotional responses or even the participants perceiving all conditions more positively due to desirability bias, which is a known limitation of vignette studies [20]. Further, all three scales used in the study do not offer the possibility of interpreting absolute values, due to their design—the 'value' of a single scale point is unknown. From the point of view of designing future personal informatics systems and, particularly, understanding if and how to build derived metrics that

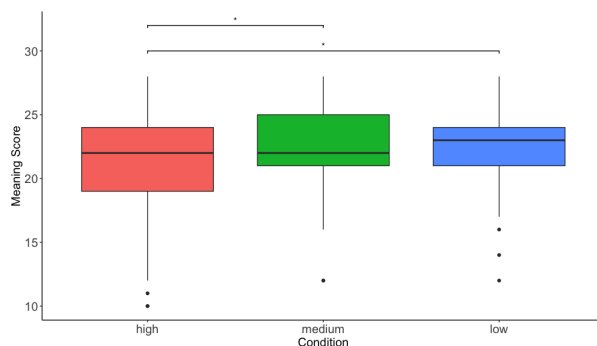


Fig. 3. Scores for perceived meaning in the health score presented in the three conditions.

offer user benefit, it is worthwhile to understand the reasons underlying the different user perceptions of the three metric abstraction levels. To that end, we conducted a qualitative study of the metric prototypes.

5 STUDY II: INTERVIEWS

Having observed that the level of abstractions influenced users' perceptions of the information the tracker demonstrates in terms of transparency, meaningfulness and reflection, we wanted to understand the reasons behind these effects. To this end, we conducted semi-structured interviews to explore users' thoughts regarding the different app screen prototypes.

5.1 Participants

We recruited $n = 12$ participants through our social networks combined with snowball sampling. One participation criterion was used for selecting participants; we included only participants who currently use, or have used a fitness tracker in the past. This criterion was used to ensure that participants know what a fitness tracker is, and have some understanding about how these devices work. The participants were aged 20–45 years, $M = 27.67$, $SD = 6.79$. Eight interviewees identified as male and four as female. All participants were residents of the European Union and were interviewed in English. We used Zoom and Microsoft Teams to conduct the interviews and record audio. Participants were asked for consent for recording before the interviews and informed that they were allowed to terminate the interview at any time. Table 1 presents details about the participants. Full demographic data, which enables further analysis, is available in the auxiliary material.

5.2 Interview Protocol

The interview started with basic demographic data collection and obtaining consent for recording. We then presented one of the three app screen prototypes (as depicted in Figure 1) and asked participants to imagine that the screen prototype presented information about their health, and to think aloud while interpreting the information. The introductory video was replaced with a narrative story in order to build rapport with the interviewer and encourage personal storytelling. The choice of the first prototype to show was counterbalanced. Based on the participants' interpretations we inquired in more detail about the meaningfulness of the information and their interpretation understanding of the health score and how it was derived. After this first inquiry, we shared all three app screen prototypes. We then asked participants to compare these three screens and to share their thoughts, interpretations and preferences. In the final part of the interview, we gave the participants the opportunity to share additional comments or thoughts and thanked them for their participation in our study.

Table 1. An overview of the participants in the interview study.

PID	Gender	Age	Country of residence	Type of tracker used	Tracking experience	First shown screen
P1	Male	23	Poland	Smart scale & Fitness tracker	several years	HIGH
P2	Male	36	Netherlands	Fitness tracker	4 years	MEDIUM
P3	Male	25	Germany	Fitness tracker	1 year	LOW
P4	Male	23	Netherlands	Fitness tracker	2 years	HIGH
P5	Male	23	Poland	Fitness tracker	2 years	MEDIUM
P6	Female	27	Germany	Fitness tracker	a few months	LOW
P7	Male	28	Germany	Fitness tracker	1-2 years	HIGH
P8	Male	27	Netherlands	Fitness tracker	1-2 years	MEDIUM
P9	Male	45	Poland	Fitness tracker	1 year	LOW
P10	Female	26	Germany	Fitness tracker	1-2 years	HIGH
P11	Female	29	Germany	Fitness tracker	1 year	MEDIUM
P12	Female	20	Poland	Smartphone application	3 years	LOW

5.3 Data Analysis

We recorded audio throughout the interviews. In total, we collected 3 hours and 37 minutes of audio. An interview lasted 18 minutes on average, $SD = 4.16$ minutes. All interviews were transcribed verbatim. We imported the transcripts into the Atlas.ti analysis software for further analysis. We applied open coding combined with thematic analysis, as described by Blandford et al. [5]. Two researchers coded a representative sample of 16% of the material. Through a set of iterative discussion rounds, we established an initial coding tree, which was then used by the first author to code the remaining material. A final discussion session was conducted to structure the coding tree and we identified three emerging themes from the material: *TRANSPARENCY*, *ACTIONABILITY*, and *METRIC DESIGN*.

5.4 Findings

Here, we present the three themes, which we constructed based on the interview data. We illustrate each theme with excerpts from the interviews.

5.4.1 Transparency. Transparency was an important aspect to all participants, and more details about the metrics were seen as valuable. A health score without any other explaining metrics (i.e. *HIGH* condition) was puzzling to all participants. None of the participants could interpret this score, and most described it as meaningless because it lacked transparency both in terms of how the derived metric was defined and how it was calculated:

To be honest, my health score looks quite concerning because the app doesn't explain what the health [metric] means, right? I don't know, maybe I am ill right now? Or the tracker diagnosed me with cancer, and it covers 25% of my body right now? So this doesn't tell me anything. (P1, *HIGH*)

Without additional metrics—examples provided included sleep, steps and stress—it was impossible for participants to interpret an abstract derived metric such as health. Similarly, this was also often the case for the *body battery* metric. Participants mentioned that they found it difficult to make sense of this metric and to conceptualise its meaning. For instance, P6 struggled to interpret the body battery. At first she figured it could be the battery level of the device, later on she realised the metric was meant as a metaphor for her personal energy level:

So, it seems like the body battery is related to the health score. Because otherwise I don't think [the user interface] would present it as part of these other ones [metrics]. Showing the device battery next to stress and health makes no sense. But then, what does this body battery mean? Is it the level

of energy – like body energy? But then how does the fitness tracker calculate that? That isn't sensor input? Does it calculate the meals that you had? The calories you consume? In that sense literally the energy that you have? Or is it based on like sleep and stress and steps, so that the more active you are the more energy you have or vice versa? Anyway, this [metric] is really confusing. It makes no sense to me. (P6, MEDIUM)

Some participants considered this abstraction and lack of transparency of metrics to be annoying, while for others it triggered curiosity. Yet, all participants indicated that they valued transparency and therefore preferred the MEDIUM and LOW conditions. Participants considered the LOW condition to be most transparent and indicated that they would like to have the option to learn more about the metrics and how they are derived. However, being offered more details about the calculation of the stress score in the LOW condition did not necessarily lead to more understanding. Several participants indicated that the MEDIUM condition offered them more than enough insight, and that more details were not needed. They concluded that there was an optimum for the degree of transparency, and that more details were not considered to be more meaningful. One participant specifically mentioned that ACTIONABILITY was more valuable than TRANSPARENCY.

5.4.2 Actionability. When participants were shown just the health metric (HIGH condition), they mentioned that they needed more information from the fitness tracker to understand how to improve their health score. The actionability of the HIGH condition was low, as described by P10:

There is stuff missing that I would really like to know. The screen gives a clear overview, it is just one graph with a percentage of 75%. But this information doesn't help me to understand anything. It is a very easy way of showing me something, but it doesn't come clear to me what to do now to actually improve my [health] score. It doesn't help me to just see the score, I need more information about what to do to improve. (P10, HIGH)

When a prototype app screen showed more metrics, participants experienced the screen as offering them more actionable insights. For instance, P1 mentioned that the HIGH condition did not offer an understanding of what to improve upon, while he could identify possible points of improvement by considering the additional metrics that the MEDIUM condition offered:

On the second screen [MEDIUM condition], you can see the details and you know that, for example, your heart rate would be too low, and this information already explains you what you can do to improve your health. As well as the stress level that would suggest that I am too stressed today. While the first screen (HIGH condition) doesn't tell me anything. (P1, MEDIUM & HIGH)

However, showing more details regarding the calculations of the metrics did not offer participants more actionable insights. For instance, knowing more about how the stress score was calculated was considered to be less valuable than a fitness tracker offering tips about how to improve the health score. One of the participants mentioned that she would like the tracker to give more tangible advice. Based on the metrics, she could try to make guesses about what to improve, yet it remained open for interpretation. She would rather see the tracker making these decisions for her, by advising her which aspect to focus on and giving tips about how to improve that aspect:

So my health score is 75%. OK, so, I am wondering what's missing right? Where do I need to improve? Do you need to sleep more? Do I need to work out more? Do I need to reduce my stress level? It's not very clear then like it's not very actionable, I would say. I mean, I like that there is an explanation [about the stress score]. But again, like the actionable part is missing. So that's not really helpful. (P6, LOW)

5.4.3 Metric Design. Notably, participants mentioned that the need for transparency and actionability differed per metric. Metrics such as heart rate, steps and to some extent sleep were considered to be much more clear and understandable than metrics such as stress, body battery and health. How steps were derived from the sensor data was never questioned, while participants were often questioning the way the device would measure stress and health. Participants also had an internalised reference for some of the metrics, which they lacked for others. For instance, one of the participants mentioned that she knows that 10.000 steps per day would be optimal and that humans need approximately 7-8 hours of sleep per night. Yet, such references were missing for other metrics such as stress or health. It remained unclear to several participants if a stress score of 70 was positive or negative, and what the start and end of the range represented. For instance:

What does [a score of] 100% for health mean? What does a score of 0% mean? What is the scale and what must be done to get [my health score up to] 100%? Is that even possible? I think such kind of information is important to me, because without knowing what the scale [of the health metric] represents, it becomes impossible to understand its meaning. (P9, HIGH)

In line with this, some participants asked if scoring a zero on their health score would mean that they were dead. This indicates that the metric was ambiguous to interpret and that participants needed more information about the range of the metrics. P10 mentioned that she considered several metrics to be more meta than others, and that this influenced her need for more information regarding their design:

The stress level will be calculated from something. So maybe from my heart rate or from relaxation phases or something? So I think the difference [between these metrics] is that, at least in my understanding, that stress level and body battery are like on a meta level. And I think that these are calculated based on several physiological data sources, they are probably combined. And that's what why I would want to know more about these [metrics] to understand what physiological data they actually use to calculate that. And I think other metrics are less calculated, maybe more [directly] measured. (P10, LOW)

When an explanation of how a particular metric was derived was unavailable, participants made assumptions when the tracker showed additional metrics; while they expressed confusion when they only saw the health score (HIGH condition), they seemed to *'fill in the blanks'* (P10) when the screen presented them with more metrics (e.g. MEDIUM or LOW condition). For instance, P11 mentioned that she understood that the health score was based on the other five metrics, which is also in line with P12's comment:

Overall health it's like, uh, when you sum up everything like the stress level and heart rate. (P12, LOW)

6 DISCUSSION

In this paper, we explored how users of fitness trackers perceive derived metrics, and inquired how the level of abstraction influences participants' perception of a health metric. We conducted two studies; an experimental vignette study and an interview study. Here, we juxtapose the results from the survey and the interview study to build an understanding of how we can effectively use derived metrics in personal informatics.

6.1 Medium Abstraction Is the Conservative Choice

In short, the two studies show that the medium level of abstraction was perceived most positively by participants, offering a moderate amount of explanation and not requiring too much speculation. We observed that the derived concept of health and other derived concepts (e.g. body battery, cf. TRANSPARENCY) prompted users to ask questions about the source of the metric. This shows that designers of future personal informatics systems

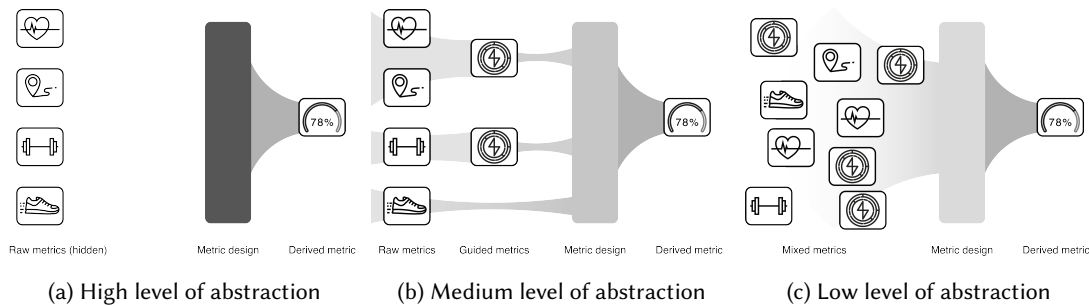


Fig. 4. A conceptual understanding of our participants' perceptions of the fitness tracker's metrics at different levels of abstraction.

need to navigate the delicate balance between fostering engagement through construal alignment (by providing derived metrics which offer a level of abstraction relevant to how users conceptualise their wellbeing, cf. [3]) and reducing meaning (and, consequently, engagement [43]). The relevance of this design question will further increase as sensors become more complex and further abstracting assessments of wellbeing from collected data may lead to a sense of loss of agency. Our study shows that prompting users to attempt to explain the source of a derived metric may be beneficial to the personal informatics experience. The design challenge that remains is to determine how much explaining is appropriate for a given metric and user. To facilitate understanding that design space, next, we propose an operational conceptualisation of a derived metric.

6.2 Towards an Anatomy of Derived metrics.

We hypothesise that the three levels elicit different responses in terms of psychological distance and thus correspond to different construal levels. Both studies demonstrated that participants related the metric to their wellbeing differently, i.e. at different levels of abstractions. When participants were shown their health information at a high level of abstraction, they were only able to create a limited interpretation and lacked transparency and meaning, which was apparent in the survey results. In contrast, the low-level metric often caused confusion, as observed in METRIC DESIGN, and triggered significantly less reflection. Figure 4 shows our conceptualisation of the three conditions.

6.2.1 High levels of abstraction obscure data. In the case of the high-level metric (Figure 4a), the design of the metric blocked users from obtaining additional information about the metric. The lack of transparency contributed to a perception of high psychological distance from the information. This, in turn, made the users unable to relate the health score to their wellbeing aspiration. In line with Niess and Wozniak [48], the lack of connection between eudaimonic aspiration and metrics leads to a sub-optimal reflection experience, which we observed.

6.2.2 Low levels weakly facilitate connections between data sources. This lack of connection was also observable in the case of the prototype on a low level of abstraction as shown in Figure 4c. While the low-level metric did not score low on transparency, participants were still unable to act upon the information provided (ACTIONABILITY). We hypothesise that the large number of metrics and thus the complexity of the interpretation involved in relating the metric to one's health was the primary reason behind its significantly lower reflection scores. In terms of the TMRM [3], users were unable to perform effective construal alignment in the low abstraction condition due to the complexity of the data presented. This result is also in line with past work which discussed how moderating complexity was a key design consideration for personal informatics [16, 49].

6.2.3 Intermediate levels of abstraction fosters reflection through interpretation. The prototype on the medium level of abstraction led to more reflection than the other two versions, yet it did not outperform the other conditions in terms of meaning or transparency. We hypothesise that the effectiveness of the medium-level metric was primarily caused by the fact that it did not produce an impression of obfuscation (like the high condition) while it also offered effective means of interpreting the data. Epstein et al. [16] showed that interpretability was a desired quality in personal informatics. In our study, the medium-level metric allowed for interpretability, thus enabling users in telling a story behind the metrics and aligning the score to their goals [48]. Not only did the medium condition offer ‘just enough’ information (TRANSPARENCY), but it also facilitated making connections between the individual metric and measurements, as we illustrate in Figure 4b. In the ACTIONABILITY theme, we saw how users selected a subset of contributing metrics at different levels of processing (e.g. steps and stress) to build their own ‘equation’ which contributed to the health score. Notably, this included metrics which were also derived but were perceived as less abstract than the health metric (we propose the term *guided metrics* for such metrics).

6.3 Designing Effective Derived Metrics

Our work shows that a derived metric can not only be subject to the technical design of how it is calculated; the presentation and, most importantly, the communicated concept behind the derived metric play a key role in how the users can reflect on the metric and how the metric can help them reach wellbeing goals. Thus, future personal informatics systems should embrace derived metrics as an additional design dimension. This is particularly relevant as all tracker metrics could be interpreted as derived or guided metrics—even heart rate measurements are interpretations of sensor signals. Thus, it may be that a true raw metric does not exist. Based on our results, we propose guidelines for such metrics.

The diversity of interpretations observed in our survey showed that personal informatics metrics are not objective entities. While sensor-based measurements can carry a sense of objectivity, all metrics are subject to interpretation and their complexity increases the users’ need for interpreting. This was particularly apparent in the TRANSPARENCY theme. Users’ understanding of the health score changed, based on the information presented. Consequently, future personal informatics systems should refrain from treating metrics as atomic outputs of the system. Instead, *a derived metric can only benefit the user if it is part of an ecology of metrics*. This finding is in line with past studies of long-term experiences of personal informatics, which found that the users’ choice of metrics is likely to evolve over time [3]. Thus, the use of a particular derived metric is dependent on a temporal and conceptual context. These contexts are likely to shift, affected by life events [29] or evolving goals [14, 48]. Well-designed derived metrics should thus provide possibilities for refinement.

Another aspect of derived metrics which was apparent in our results was how participants eagerly made causal connections between the metrics. We observed how every metric, and particularly a derived metric, was accompanied by a story. In the ACTIONABILITY theme, we observed how connections between different metrics facilitated reflection. Thus, the users desired to understand the background of the metric and its connections to specific sequences of their wellbeing-related actions. The medium-level metric obtained the highest reflection score in the survey and the users were most eager to interpret it in the interviews. This suggests that it is a key design quality for derived metrics to *support effective storytelling through providing the opportunity to connect different metrics into stories*.

Our work indicates that the medium-level prototype was most effective in fostering reflection as it featured a balanced mix of different metrics contributing to the health metric. Users were able to effectively choose not only their interpretation of the metric, but also the preferred level of abstraction of that interpretation. Thus, the health score was more or less abstract to different users, according to their preference. This showcases how there is a need to design for ‘the middle way’ in derived metrics—embracing a balance in the complexity and

abstractness. This was illustrated both in the fact that the medium version in our study prompted most reflection and through qualitative insights in the METRIC DESIGN theme. In future personal informatics systems, there is a need to *provide a mix of raw and guided metrics which contribute to the derived metrics to allow the users to adjust their interpretation of the derived metric to their desired level of abstraction.*

We note that, in our study, we examined the three abstraction levels as mutually exclusive, in order to understand the differences between the three conditions. This is not the case in most commercial tracker interfaces which often follow the ‘details on demand’ visualisation mantra [52], particularly in terms of visualising metrics over time. Arguably, traditional fitness measurement supports only *revisiting* [6], while derived metrics may support *revisiting* and *explanation*. Thus, the design solution for the active user who regularly interprets data using the tracker tool is making multiple abstraction levels available on the demand, facilitating multiple means of supporting reflection. However, as most personal informatics interactions are glances [9], choosing the default abstraction level remains a key design choice.

Finally, our study showed that the current state of design metrics provided in commercial trackers—with little to no explanation—lacked transparency and made it impossible to fully use the tracker’s potential for fostering reflection. This was evident both in the quantitative data and in the METRIC DESIGN theme. Yet, full transparency was also not a desired design approach as it led to a loss of reflection potential and limited actionability. Consequently, the design of future derived metrics should apply the ‘middle way’ principle. Future trackers should *reveal how a derived metric is obtained and calculated to the user to a certain degree.* The metric should allow the user to understand the component contributing to the metric, while leaving ample room for interpretation.

6.4 Limitations

We recognise that there are certain limitations to our study. First, we observe that the participants in our study were all from a Western background. While this choice is in line with other personal informatics studies, it limits the applicability of the work. However, focusing on a Western population allowed us to conduct a focused experiment and reduce confounding variables such as that the conceptualisation of health and the interpretation of derived metrics varies among cultures [31]. Thus, there is a need to further study derived metrics across cultures as habits and motivations for fitness trackers are a diverse set of practices [47]. Moreover, we acknowledge that our inquiry focused on fitness tracker metrics and we do not know how our findings generalise to other domains of personal informatics.

We note that there are, to the best of our knowledge, no validated scales available that allow us to assess how CLT applies to experiences in the digital domain. While our analysis suggests that CLT can be applied to explain the results of our work, there is a need for developing further psychometric instruments to fully confirm our interpretation with statistical analysis. The choice of measures used in our studies also limits the scope of our inquiry. Based on past work, we considered transparency, reflection and meaning as key dimensions in a positive personal informatics experience. We do, however, recognise that other facets of personal tracking should also be explored in future research.

Further, we remark that the use of fictional data was a study design choice which may have influenced the findings of our study. Our approach is contrary to research efforts which postulated personal engagement with data in order to understand personal informatics, e.g. [44]. Our choice of study design was motivated by the need to mitigate possible negative consequences of participating in the study [12, 57] While we recognise that using real data would have resulted in participant responses of increased ecological validity, it would also render a comparison between the conditions impossible. Past work indicates that users interpret data in the context of their own goals, life context and personal models of performance [3, 48]. Consequently, the subjective assessment of the data underlying the presented prototype would be the primary factor determining the users’ perception of

the prototype. Further, engaging with one's own data could result in discovery [44] for a subset of the participants, leading to a sample imbalance. Given that these confounding factors cannot be reliably measured, we would be unable to control for them, which would, in turn, severely impact the validity of the comparisons in our studies. In the quantitative results, we can observe that the effect sizes are low, which may be due to the use of fictional data. Future work should explore how our findings can be applied to studies with real data while avoiding negative consequences of changes in data presentation, possibly through multiple levels of abstraction available at the same time.

In order to avoid familiarity bias, in the form of the users being acquainted with some commercially available derived metrics, we chose to use a fictional health metric in our study. This metric, to the best of our knowledge, is not available in commercial fitness trackers. Yet, we recognise that a new metric may require extra effort from the users and elicit different responses than a familiar product. Interestingly, the qualitative data does not show that the participants were opposed to the notion of a health metric per se. We hope that future research can juxtapose our results with studies of existing derived metrics to understand the interdependence between how a derived metric is presented and its relation to everyday experience.

Finally, it is of importance to remark that, as any study which uses a design artefact, the results of this inquiry are influenced by the speculative interface designs which we used to elicit feedback from users. While we aimed to work with designs that closely resemble commercial fitness trackers to increase ecological validity, there is inherent bias in any design. Thus, it remains a challenge for future research to validate the insights proposed here across a range of personal informatics designs.

7 CONCLUSION

This paper explored how users of fitness trackers perceive derived metrics, and how such metrics can support users' wellbeing. We conducted a two-step study which consisted of an online survey and an interview study. We examined three ways of presenting a derived metric for health at three levels of abstraction. Our findings demonstrated that a medium level of abstraction led to the highest potential for reflection. Our analysis suggests that this was due to the fact that a medium level of abstraction allowed for construal alignment and supported participants in identifying underlying reasons for the metric score. In addition, our work shows that derived metrics may provide additional opportunities for reflection, but their complexity increases the users' need for interpretation. This indicates that derived metrics form an additional design dimension for personal informatics systems. This paper offers four guidelines for the design of such metrics. We hope that our work contributes to an understanding of users' perception of derived metrics and will inspire the design of effective derived metrics for future personal informatics systems.

REFERENCES

- [1] Eric P. S. Baumer, Mark Blythe, and Theresa Jean Tanenbaum. 2020. *Evaluating Design Fiction: The Right Tool for the Job*. Association for Computing Machinery, New York, NY, USA, 1901–1913. <https://doi.org/10.1145/3357236.3395464>
- [2] Marit Bentvelzen, Jasmin Niess, Mikołaj P. Woźniak, and Paweł W. Woźniak. 2021. The Development and Validation of the Technology-Supported Reflection Inventory. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 366. Association for Computing Machinery, New York, NY, USA, 1–8. <https://doi.org/10.1145/3411764.3445673>
- [3] Marit Bentvelzen, Jasmin Niess, and Paweł W. Woźniak. 2021. The Technology-Mediated Reflection Model: Barriers and Assistance in Data-Driven Reflection. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 246. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3411764.3445505>
- [4] Johannes Bircher. 2005. Towards a Dynamic Definition of Health and Disease. *Medicine, Health Care and Philosophy* 8, 3 (Nov. 2005), 335–341. <https://doi.org/10.1007/s11019-005-0538-y>
- [5] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI Research: Going Behind the Scenes. *Synthesis Lectures on Human-Centered Informatics* 9 (2016), 1–115. <https://doi.org/10.2200/S00706ED1V01Y201602HCI034>
- [6] Janghee Cho, Tian Xu, Abigail Zimmermann-Niefield, and Stephen Voida. 2022. Reflection in Theory and Reflection in Practice: An Exploration of the Gaps in Reflection Support among Personal Informatics Apps. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. Number 366. Association for Computing Machinery, New York, NY, USA, 1–8. <https://doi.org/10.1145/3511764.3545673>

- Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 142, 23 pages. <https://doi.org/10.1145/3491102.3501991>
- [7] Sunny Consolvo, David W McDonald, Tammy Toscos, Mike Y Chen, Jon Froehlich, Beverly Harrison, Predrag Klasnja, Anthony LaMarca, Louis LeGrand, Ryan Libby, Ian Smith, and James A Landay. 2008. Activity Sensing in the Wild: A Field Trial of UbiFit Garden. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1797–1806. <https://doi.org/10.1145/1357054.1357335>
- [8] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5 (Aug. 2008), 455. <https://doi.org/10.1007/s11257-008-9051-3>
- [9] Andrea Cuttone, Michael Kai Petersen, and Jakob Eg Larsen. 2014. Four data visualization heuristics to facilitate reflection in personal informatics. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 541–552.
- [10] Nediya Daskalova, Jina Yoon, Yibing Wang, Cintia Araujo, Guillermo Beltran, Nicole Nugent, John McGeary, Joseph Jay Williams, and Jeff Huang. 2020. SleepBandits: Guided Flexible Self-Experiments for Sleep. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376584>
- [11] Xianghua (Sharon) Ding, Shuhan Wei, Xinning Gui, Ning Gu, and Peng Zhang. 2021. Data Engagement Reconsidered: A Study of Automatic Stress Tracking Technology in Use. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Number 535. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3411764.3445763>
- [12] Elizabeth Victoria Eikey, Clara Marques Caldeira, Mayara Costa Figueiredo, Yunan Chen, Jessica L. Borelli, Melissa Mazmanian, and Kai Zheng. 2021. Beyond Self-Reflection: Introducing the Concept of Rumination in Personal Informatics. *Personal Ubiquitous Comput.* 25, 3 (jun 2021), 601–616. <https://doi.org/10.1007/s00779-021-01573-w>
- [13] Elizabeth V Eikey and Madhu C Reddy. 2017. "It's Definitely Been a Journey": A Qualitative Study on How Women with Eating Disorders Use Weight Loss Apps. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 642–654. <https://doi.org/10.1145/3025453.3025591>
- [14] Tina Ekhtiar, Rúben Gouveia, Armağan Karahanoglu, and Geke Ludden. 2022. Reflection during goal setting: An analysis of popular personal informatics apps. (2022).
- [15] Chris Elsdén, David S. Kirk, and Abigail C. Durrant. 2016. A Quantified Past: Toward Design for Remembering With Personal Informatics. *Human-Computer Interaction* 31, 6 (Nov. 2016), 518–557. <https://doi.org/10.1080/07370024.2015.1093422> Publisher: Taylor & Francis _eprint: <https://doi.org/10.1080/07370024.2015.1093422>.
- [16] Daniel Epstein, Felicia Cordeiro, Elizabeth Bales, James Fogarty, and Sean Munson. 2014. Taming data complexity in lifelogs: exploring visual cuts of personal informatics data. In *Proceedings of the 2014 conference on Designing interactive systems (DIS '14)*. Association for Computing Machinery, New York, NY, USA, 667–676. <https://doi.org/10.1145/2598510.2598558>
- [17] Daniel A. Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M. Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, Payam Dowlatyari, Craig Hilby, Sazedra Sultana, Elizabeth V. Eikey, and Yunan Chen. 2020. Mapping and Taking Stock of the Personal Informatics Literature. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 126 (Dec. 2020), 38 pages. <https://doi.org/10.1145/3432231>
- [18] Daniel A. Epstein, An Ping, James Fogarty, and Sean A. Munson. 2015. A lived informatics model of personal informatics. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 731–742. <https://doi.org/10.1145/2750858.2804250>
- [19] Sergio Felipe, Aneesh Singh, Caroline Bradley, Amanda CdeC Williams, and Nadia Bianchi-Berthouze. 2015. Roles for personal informatics in chronic pain. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '15)*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, 161–168.
- [20] Maria F Fernandes and Donna M Randall. 1992. The nature of social desirability response effects in ethics research. *Business Ethics Quarterly* (1992), 183–205.
- [21] Janet Finch. 1987. The Vignette Technique in Survey Research. *Sociology* 21, 1 (Feb. 1987), 105–114. <https://doi.org/10.1177/0038038587021001008> Publisher: SAGE Publications Ltd.
- [22] Fitbit. 2021. Stress Management - Stress Watch & Monitoring | Fitbit. <https://www.fitbit.com/global/us/technology/stress>
- [23] Fitbit. 2021. Stress Management - Stress Watch & Monitoring | Fitbit. <https://www.fitbit.com/global/us/technology/stress#premium>
- [24] Fitbit. 2021. What's sleep score in the Fitbit app? https://help.fitbit.com/articles/en_US/Help_article/2439.htm
- [25] Rowanne Fleck and Geraldine Fitzpatrick. 2010. Reflecting on reflection: framing a design landscape. In *Proceedings of the 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction (OZCHI '10)*. Association for Computing Machinery, New York, NY, USA, 216–223. <https://doi.org/10.1145/1952222.1952269>
- [26] Garmin. 2021. Body Battery Frequently Asked Questions | Garmin Support. <https://support.garmin.com/en-US/?faq=VOFJAsiXut9K19k1qEn5W5>

- [27] Garmin. 2021. Garmin Body Battery: How Does It Work? Everything You Need to Know. <http://runningwithrock.com/garmin-body-battery/>
- [28] Garmin. 2021. What Is the Stress Level Feature on My Garmin Watch? | Garmin Support. <https://support.garmin.com/en-US/?faq=WT9BmhjacO4ZpxbCc0EKn9>
- [29] Rebecca Gulotta, Jodi Forlizzi, Rayoung Yang, and Mark Wah Newman. 2016. Fostering Engagement with Personal Informatics Systems. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16)*. ACM, New York, NY, USA, 286–300. <https://doi.org/10.1145/2901790.2901803>
- [30] Teng Han, Xiang Xiao, Lanfei Shi, John Canny, and Jingtao Wang. 2015. Balancing Accuracy and Fun: Designing Camera Based Mobile Games for Implicit Heart Rate Monitoring. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 847–856. <https://doi.org/10.1145/2702123.2702502>
- [31] Cecil Helman. 2007. *Culture, Health and Illness, Fifth edition*. CRC Press. Google-Books-ID: uz59BgAAQBAJ.
- [32] Dandan Huang, Melanie Tory, Bon Adriel Aseniero, Lyn Bartram, Scott Bateman, Sheelagh Carpendale, Anthony Tang, and Robert Woodbury. 2015. Personal Visualization and Personal Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* 21, 3 (March 2015), 420–433. <https://doi.org/10.1109/TVCG.2014.2359887> Conference Name: IEEE Transactions on Visualization and Computer Graphics.
- [33] Veronika Huta and Richard M. Ryan. 2010. Pursuing Pleasure or Virtue: The Differential and Overlapping Well-Being Benefits of Hedonic and Eudaimonic Motives. *Journal of Happiness Studies* 11, 6 (Dec. 2010), 735–762. <https://doi.org/10.1007/s10902-009-9171-4>
- [34] Matthew Kay, Eun Kyoung Choe, Jesse Shepherd, Benjamin Greenstein, Nathaniel Watson, Sunny Consolvo, and Julie A. Kientz. 2012. Lullaby: A Capture & Access System for Understanding the Sleep Environment. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (Pittsburgh, Pennsylvania) (UbiComp '12)*. Association for Computing Machinery, New York, NY, USA, 226–234. <https://doi.org/10.1145/2370216.2370253>
- [35] Amrith Krishna, Madhumita Mallick, and Bivas Mitra. 2016. SleepSensei: An Automated Sleep Quality Monitor and Sleep Duration Estimator. In *Proceedings of the First Workshop on IoT-Enabled Healthcare and Wellness Technologies and Systems (IoT of Health '16)*. Association for Computing Machinery, New York, NY, USA, 29–34. <https://doi.org/10.1145/2933566.2933570>
- [36] Ian Li, Anind Dey, and Jodi Forlizzi. 2010. A Stage-Based Model of Personal Informatics Systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 557–566. <https://doi.org/10.1145/1753326.1753409>
- [37] Zilu Liang and Bernd Ploderer. 2020. How does fitbit measure brainwaves?: A qualitative study into the credibility of sleep-tracking technologies. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1 (2020), 1–29. <https://doi.org/10.1145/3380994>
- [38] Zilu Liang, Bernd Ploderer, and Mario Alberto Chapa-Martell. 2017. Is Fitbit Fit for Sleep-Tracking? Sources of Measurement Errors and Proposed Countermeasures. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 476–479. <https://doi.org/10.1145/3154862.3154897>
- [39] James J Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry B Strub. 2006. Fish'N'Steps: Encouraging Physical Activity with an Interactive Computer Game. In *Proceedings of the 8th International Conference on Ubiquitous Computing (UbiComp'06)*. Springer-Verlag, Berlin, Heidelberg, 261–278. https://doi.org/10.1007/11853565_{ }16
- [40] Fannie Liu, Laura Dabbish, and Geoff Kaufman. 2017. Supporting Social Interactions with an Expressive Heart Rate Sharing Application. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (9 2017). <https://doi.org/10.1145/3130943>
- [41] Ian McDowell et al. 2006. *Measuring health: a guide to rating scales and questionnaires*. Oxford University Press, USA.
- [42] D. Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. Trust in a specific technology: An investigation of its components and measures. *ACM Transactions on Management Information Systems* 2, 2 (July 2011), 12:1–12:25. <https://doi.org/10.1145/1985347.1985353>
- [43] Elisa D. Mekler and Kasper Hornbæk. 2016. Momentary Pleasure or Lasting Meaning? Distinguishing Eudaimonic and Hedonic User Experiences. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 4509–4520. <https://doi.org/10.1145/2858036.2858225>
- [44] Jimmy Moore, Pascal Goffin, Jason Wiese, and Miriah Meyer. 2022. An Interview Method for Engaging Personal Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 4, Article 173 (dec 2022), 28 pages. <https://doi.org/10.1145/3494964>
- [45] Andre Müller, Nan Xin Wang, Jiali Yao, Chuen, Chuen Seng Tan, Ivan Cherh, Chiet Low, Nicole Lim, Agnes Tan, and Falk Mueller-Riemenschneider. 2019. Heart Rate Measures From Wrist-Worn Activity Trackers in a Laboratory and Free-Living Setting: Validation Study. *Journal of Medical Internet Research* (2019). <https://doi.org/10.2196/14120a>
- [46] Jasmin Niess, Kristina Knaving, Alina Kolb, and Pawel W. Woźniak. 2020. Exploring Fitness Tracker Visualisations to Avoid Rumination. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.org/10.1145/3379503.3405662>
- [47] Jasmin Niess, Pawel W. Woźniak, Yomna Abdelrahman, Passant ElAgroudy, Yasmeeen Abdrabou, Caroline Eckerth, Sarah Diefenbach, and Kristina Knaving. 2021. 'I Don't Need a Goal': Attitudes and Practices in Fitness Tracking beyond WEIRD User Groups. Association for

- Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/3447526.3472062>
- [48] Jasmin Niess and Paweł W. Woźniak. 2018. Supporting Meaningful Personal Fitness: the Tracker Goal Evolution Model. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173745>
- [49] Laura R. Pina, Sang-Wha Sien, Teresa Ward, Jason C. Yip, Sean A. Munson, James Fogarty, and Julie A. Kientz. 2017. From Personal Informatics to Family Informatics: Understanding Family Practices around Health Monitoring. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2300–2315. <https://doi.org/10.1145/2998181.2998362>
- [50] 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. <https://doi.org/10.1145/3196829>
- [51] Kathleen Ryan, Conor Linehan, and Samantha Dockray. 2021. Appropriation of Digital Tracking Tools in an Online Weight Loss Community: Individual and Shared Experiences. In *Designing Interactive Systems Conference 2021*. Association for Computing Machinery, New York, NY, USA, 999–1014. <https://doi.org/10.1145/3461778.3462092>
- [52] Ben Shneiderman. 2003. The eyes have it: A task by data type taxonomy for information visualizations. In *The craft of information visualization*. Elsevier, 364–371.
- [53] Paul D. Trapnell and Jennifer D. Campbell. 1999. Private self-consciousness and the five-factor model of personality: Distinguishing rumination from reflection. *Journal of Personality and Social Psychology* 76, 2 (1999), 284–304. <https://doi.org/10.1037/0022-3514.76.2.284> Place: US Publisher: American Psychological Association.
- [54] Yaacov Trope and Nira Liberman. 2010. Construal-Level Theory of Psychological Distance. *Psychological Review* 117, 2 (2010), 440–463. <https://doi.org/10.1037/a0018963>
- [55] Gregorio Villalobos, Rana Almaghrabi, Behnoosh Hariri, and Shervin Shirmohammadi. 2011. A Personal Assistive System for Nutrient Intake Monitoring. In *Proceedings of the 2011 International ACM Workshop on Ubiquitous Meta User Interfaces (Ubi-MUI '11)*. Association for Computing Machinery, New York, NY, USA, 17–22. <https://doi.org/10.1145/2072652.2072657>
- [56] WHOOP. 2021. WHOOP Experience: Recovery, Strain and Sleep Metrics Optimize Training. <https://www.whoop.com/experience/>
- [57] Paweł W. Woźniak, Przemysław Piotr Kucharski, Maartje M.A. de Graaf, and Jasmin Niess. 2020. Exploring Understandable Algorithms to Suggest Fitness Tracker Goals that Foster Commitment. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society (NordiCHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3419249.3420131>