Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity

RAFAL KOCIELNIK, Human Centered Design & Engineering, University of Washington, USA
LILLIAN XIAO, Human Centered Design & Engineering, University of Washington, USA
DANIEL AVRAHAMI, FXPAL, USA
GARY HSIEH, Human Centered Design & Engineering, University of Washington, USA

Mobile, wearable and other connected devices allow people to collect and explore large amounts of data about their own activities, behavior, and well-being. Yet, learning from-, and acting upon such data remain a challenge. The process of reflection has been identified as a key component of such learning. However, most tools do not explicitly design for reflection, carrying an implicit assumption that providing access to self-tracking data is sufficient. In this paper, we present *Reflection Companion*, a mobile conversational system that supports engaging reflection on personal sensed data, specifically physical activity data collected with fitness trackers. *Reflection Companion* delivers daily adaptive mini-dialogues and graphs to users' mobile phones to promote reflection. To generate our system's mini dialogues, we conducted a set of workshops with fitness trackers users, producing a diverse corpus of 275 reflection questions synthesized into a set of 25 reflection mini dialogues. In a 2-week field deployment with 33 active Fitbit users, we examined our system's ability to engage users in reflection through dialog. Results suggest that the mini-dialogues were successful in triggering reflection and that this reflection led to increased motivation, empowerment, and adoption of new behaviors. As a strong indicator of our system's value, 16 of the 33 participants elected to continue using the system for two additional weeks without compensation. We present our findings and describe implications for the design of technology-supported dialog systems for reflection on data.

CCS Concepts: • **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; *Empirical studies in HCI*; Field studies; Natural language interfaces

Additional Key Words and Phrases: Conversational design; reflection; conversational AI; self-learning; behavior change; physical activity; reflective dialogues

ACM Reference Format:

Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol* 2, 2, Article 70 (June 2018), 26 pages. https:/doi.org/10.1145/3214273

1 INTRODUCTION

Advancements in sensing and communication technologies have enabled the introduction of connected sensing into our everyday lives. Mobile, wearable, and IoT consumer devices allow people to collect and

Authors' addresses: Rafal Kocielnik, Human Centered Design & Engineering, University of Washington, Seattle, WA, 98195, USA, rkoc@uw.edu; Lillian Xiao, Human Centered Design & Engineering, University of Washington, Seattle, WA, 98195, USA, lillianx@uw.edu; Daniel Avrahami, FXPAL, Palo Alto, CA, 94304, USA, daniel.avrahami@gmail.com; Gary Hsieh, Human Centered Design & Engineering, University of Washington, Seattle, WA, 98195, USA, garyhs@uw.edu

2474-9567/2018/6–ART70 $15.00
https://doi.org/10.1145/3214273

Fig. 1. Example of an actual user exchanges with our system's mini-dialogues on the left. On the right a block diagram of an example dynamic mini dialogue with: actual user replies, user intents recognized based on free-text replies, and the system tailored follow-ups. The red boxes represent a path where user reply was not recognized and has been handled by a "generic" (non-tailored) follow-up.

examine large amounts of data about their activities, behavior, and wellbeing. However, a gap remains between our ability to collect and visualize data, and our ability to learn from-, and act upon this data in meaningful ways [55]. A key component for bridging this gap is to facilitate reflection [7, 17, 56]. The value of engaging users in reflection has been identified as a key element of successful health behavior change [56, 59] and is an important step in stage-based models of personal informatics [27, 54]. Through the process of reflection, users can increase their self-knowledge [8], formulate realistic behavior change goals [53], and increase self-control while promoting positive behaviors [56]. Reflection has been considered an impetus that moves the individual from examinations of his or her data to action [7].

Despite the importance of reflection, personal informatics models reveal little about how reflection can, or should be triggered [8]. Consequently technology has struggled to successfully support refection in practice [29, 68]. As noted in [8] *"prior work carries an implicit assumption that by providing access to data that has been 'prepared, combined, and transformed' for the purpose of reflection, reflection will occur."* Indeed, one main means of facilitating reflection in behavior change and personal informatics relies on visualizations of self-tracking data, such as Fish'n'Steps [57], UbiFitGarden [22] for physical activity; Affect Aura [60] for affective states and LifelogExplorer [47] for stress. The other approach relies on journaling [65], such as SleepTight [17] for sleep and Affective Diary [58] for manual journaling of emotions. Both of these approaches assume that reflection will occur naturally when data is presented. However, reflection is time consuming and not necessarily something that comes naturally to people [29]. In many cases people need a reason to reflect or at least an encouragement to do so [36, 62]. Results from our exploratory workshop with 12 active Fitbit users further corroborate these findings, revealing that such users engage in none or very limited reflection. Also, with existing tools, they find it boring, repetitive, and sometimes even demotivating. How should systems facilitate reflection on self-tracked data? Further, can a conversational system support reflection that is engaging rather than burdensome?

Based on the domain of personal counseling, supporting reflection through conversation seems like a promising approach. Several personal counseling techniques, such as motivational interviewing [74] and commercial behavior change programs (e.g., Weight Watchers [40]) rely on engaging and insightful

conversations with the goal of triggering reflection on one's own activity. Personal coaches *"repeatedly ask questions to get at hidden motivations"* and that asking reflection questions can help people understand and articulate their underlying needs and goals [53]. Such conversations can elicit contemplative [41] and metacognitive [28] thinking, encouraging people to think about the needs and wants beyond their first answers that come to mind.

In this work, we explore the feasibility of using conversational mini-dialogues for triggering reflection on physical activity data. We are motivated by the fact that despite the popularity of commercial fitness tracking tools such as Fitbit and Garmin, many of these tools do not currently explicitly support reflection [17]. Thus, to explore how reflection might be integrated into these tools, we designed and deployed a conversational system (see Figure 1) that delivers reflection prompts on 3 levels based on learning theory: *Noticing*, *Understanding*, and *Future Actions* [62]. Our system delivers daily adaptive mini-dialogues along with graphs of the user's data over MMS to the user's mobile device. To generate the system's mini-dialogue flows, we conducted a series of workshops with 12 active Fitbit users. From these workshops, a set of 275 prompts were generated and later synthesized to form 25 mini-dialogues used by our system.

Using a 2-week long field study in which 33 Fitbit users received one mini-dialogue a day, we demonstrate that our system is able to successfully trigger engaging reflection that in turn can lead to an increase in mindfulness, motivation, adoption of new behaviors around physical activity, and empowerment through increased understanding of barriers and formulation of concrete action plans. Further, we identify how different aspects of our conversation design deepen reflection by making it more actionable, personal, and accountable. We show that follow-up prompts are most useful when they dynamically build upon user responses. As a strong indication of success in creating an engaging and valuable system, 16 of the 33 participants elected to continue using the system for two additional weeks without compensation.

This work makes the following contributions: 1) We adopt a structured reflection model to inform the design of conversational mini-dialogues for proactively supporting user learning from physical activity data collected by wearable trackers; 2) We present a mobile conversational system demonstrating that our approach is feasible, beneficial, and appreciated by the users; and 3) We describe a participatory process of generating diverse mini-dialogues and offer design insights that can inform future design in conversational reflection.

2 RELATED WORK

The abundance of wearable sensors, mobile and IoT devices, as well as other types of UbiComp technologies together with connected cloud platforms enable collection and storage of copious amounts of data [31, 50]. This data can be related to the personal physical activity [35], stress monitoring [48, 49], urban spaces [30], all the way to energy conservation [25]. All of this data, however, becomes truly useful only when it facilitates learning for the purpose of increasing awareness of one's own behavior or the environment [15], making better future decisions [26], or supporting increased understanding [19]. One of the key ways of supporting self-learning from such data collected by UbiComp technologies is through the process of reflection [55].

In fact, a stage-based model of personal informatics states that *"collection and reflection are the core aspects of every personal informatics system"* [54]. Yet, a gap exists in understanding how the process of reflection can be supported through technology [73]. Indeed, designing for reflection is still in its infancy [7, 29]. As a result, much HCI research that attempts to inform design for reflection is based on structured reflection models from learning theory where such models and theories are more mature [29]. Thus far, such works have shown how learning models can be adapted to HCI for analyzing reflection [29], reviewing it [8], or designing for it on a conceptual level [68]. Critically, little work has been done in supporting structured reflection in deployed behavior-change systems. Our work aims to fill this gap.

2.1 Structured Reflection Models

A number of theoretical works on reflection have been developed in learning sciences. Some of them focus on exploring the nature of reflection itself [13] or the place of reflection in different professions [62, 69] without

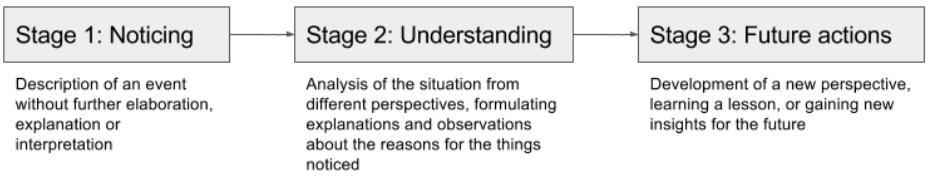

Fig. 2. Reflection depicted as a process with stages of levels synthesized based on multiple structured reflection models.

necessarily focusing specifically on how reflection could be triggered. Nevertheless, a number of works discuss the *right environment* and *conditions* for reflection [36]. As discussed in [62], reflection takes time and therefore creating and allowing time for reflection is essential. On top of that, reflection is seen by many as a developmental process that can be supported to help people become more reflective over time [36, 62, 80]. When the purpose of reflection is more formal, such as in the case of learning from one's physical activity patterns, structured support or guidance around reflection is of particular value [29]. In such applications, structured reflection models provide insights for designing reflection-centered interactions and offer support for how reflection can be supported to evolve with time [29]. Such models see reflection as a process with stages or levels. Kolb's learning cycle defines 4 stages [51], Gibb's reflective cycle proposes 6 stages [34], and both Moon's levels of learning [62] and Bain's 5Rs framework [5] suggest 5 stages (or levels) of reflection. Fleck and Fitzpatrick have proposed an adaptation of these models to HCI, defining 5 levels of reflection [29]. These models, however, were developed to analyze reflection post factum and may be too fine-grained for direct use in design. Conveniently, Atkins and Murphy [4] in their review of literature on reflection, identified three commonly-shared stages: *awareness of uncomfortable feeling and thought*, *critical analysis*, and *development of new perspective*. Our approach aligns with the three stages from Atkins and Murphy, renaming them for simplicity into: *Noticing*, *Understanding*, and *Future actions* (Figure 2).

Stage 1 – Noticing This stage focuses on building awareness of events and behavior patterns, but without an explicit attempt at explaining or understanding reasons. The stage is aligned with Fleck and Fitzpatrick's *revisiting* and *reflective description* levels where description of an event is provided without further elaboration, explanation, or interpretation.

Stage 2 – Understanding This stage focuses on analysis of the situation from different perspectives, formulating explanations and observations about the reasons for the things noticed. The stage is aligned with Fleck and Fitzpatrick's *dialogic reflection* level where cycles of interpreting and questioning as well consideration of different explanations, hypotheses, and viewpoints are taking place.

Stage 3 – Future Action In this stage, Understanding leads to development of a new perspective, learning a lesson, or gaining new insights for the future. In terms of Fleck and Fitzpatrick's levels, this step aligns with levels of *transformative reflection* and *critical reflection* where past events are revisited with intent to re-organize them and do something differently in the future. Personal assumptions are challenged, leading to change in practice and understanding. Here also wider implications of actions are taken into consideration.

2.2 Conversational Approach towards Reflection

As pointed out in [8], many current approaches in personal informatics support reflection as an activity of *"looking at lists of collected personal information or exploring or interacting with information visualization"*. Not until very recently, have researchers started to even study the reflection questions users themselves may have when exploring their personal data [19]. Yet human coaches of behavior change take much more active approaches, asking reflective questions that can help people articulate their underlying needs and goals and increase their motivation [53]. In one example, people who were asked to think about why they eat snacks before making a choice were more likely to choose healthy options [32]. In fact, research has shown that asking people their reasons for doing an activity triggers underlying motivations and leads to focus on higher-

level goals [14, 76]. Specifically asking 'why' questions twice has been shown to be effective [29] as well as asking people to take more time to think about the question and to write longer answers [32].

The paradigm of computers as social actors [63] argues that people will apply social rules to a computer. This suggests that successful human counseling techniques might also work effectively in computer based delivery. Indeed an accumulated body of research has demonstrated the efficacy of human-counseling-inspired computer-based interventions [9, 44]. In fact some research suggests that in computer-based counseling services, without dialogues, people may be less inclined to comply and may provide only superficial answers to questions [53]. Consequently, in virtual-agent-related research, it has been a goal to construct an engaging, long-term relationship with the user [12].

A comprehensive review by Bickmore and Giorgino on work in health education and behavior change dialogue systems [10] has revealed application domain spanning exercise, diet, smoking cessation promotion, medication adherence, and chronic disease management education and promotions. Most common approaches relied on building prescriptive and persuasive dialogues that would tell the user what to do rather than guide the user to reflect and explore their own goals and motivations [1]. Reflection was not the main focus, and when used related to reflective listening from motivational interviewing, where the approach for reflection is to provide a "tell me more" prompt or simply restate what the user said [64]. While such approach may work for one-time interactions, it quickly becomes repetitive and boring in long-term repeated interactions [12]. Indeed, most of the approaches are not designed for long-term repeated interaction with one user, but rather one-time interactions with many different users. Revisiting such dialogue again provides the same fixed interaction. This could be a problem in personal informatics when new data is collected daily and frequent, but short, reflective sessions could be appropriate. In fact, in the FitTrack study [12], several subjects mentioned that repetitiveness in the system's dialog content was responsible for their losing motivation to continue working with the system and following recommendations. In general although counseling interventions delivered by computer have been found effective, high drop-out rates due to low user engagement during interaction limit their long-term adoption and potential impact [66].

Reflection in personal informatics offers a unique challenge for computer-based dialogue systems as it relies on open user responses, is relatively frequent and requires novel perspectives to engage users each time over long-term [45]. Nevertheless, despite recent advances in ML/NLP techniques, designing a feasible conversational system that can handle the freedom of user expression needed for the purpose of reflection and engage users for a long-term needed for reflection to occur is still a challenging problem [11]. Aspects of properly handling unexpected user responses and potential misrecognitions are challenging [70], but might potentially be mitigated by appropriate design.

3 WORKSHOPS FOR DESIGNING REFLECTION QUESTIONS

A critical component of a conversational system for reflection is the questions that solicit reflection from the users. In this work, we employed a workshop-based approach to develop these reflection prompts. Working with 12 existing users of activity trackers, the workshop approach helped generated a diverse set of reflection prompts. It also enabled us to understand current (limited) reflection practices, and offered valuable insights on how to design the conversational system.

3.1 Procedure

We recruited participants through social media posts and mailing lists. Prior to the workshop, participants were asked to share their daily, weekly, and long-term behavior change goals, as well as take screenshots of their self-tracking data from the previous 2 weeks, including measures of steps, calories, distance traveled, floors, resting heart rate, and sleep, if available.

We started each session with a semi-structured discussion around reflection, in which we asked 5 questions: 1) What reflection on behavior change have you engaged in? 2) What is the goal of reflection for

Table 1. Examples of reflective questions generated during the workshop sessions. Questions are grouped by the main prompted categories (rows) and categories identified in through affinity diagramming (columns). Only the 6 most frequent categories are shown. The five white cells represent intersections for which the workshop participants generated no questions. For creating diverse and novel questions, we suggested questions for these intersections ourselves.

	General/ Context (n=27)	Goals (n=50)	Tracking (n=29)	Observations/ Patterns (n=69)	Motivations (n=20)	Plans/ Scheduling (n=24)
Noticing (n=76)	What are you doing to be more active?	How many days did you meet your goals?	How many days did you wear your tracker this week?	Which days do you walk more?	Did you notice any especially motivating moments this week?	What were the discrepancies between your plans and your actual activities?
Understanding (n=116)	What are the top 3 reasons you're stationary?	Why do you only hit your goal on certain days?	What actions lead you to logging your food?	What happened during peaks/low points during your week?	Why were you sometimes unmotivated this week?	Why did you skip some part of your plan this week?
Future Actions (n=83)	What events could affect your activity next week?	Should you reevaluate your future goals?	What other metrics do you want to track?	How can you avoid low activity days next week?	How can you encourage yourself to exercise regularly?	How can you set yourself up to have a day similar to successful days before?

you? 3) How often/when do you usually reflect? 4) Where do you access your self-tracking data? 5) What are the main challenges in reflecting for you? For each question, participants wrote their responses on post-its, shared each response with the group, and provided clarification when needed. Each session lasted an hour on average. After the discussion, each participant was given a paper form with instructions to write at least two reflective questions for each of the 3 reflective categories of *Noticing*, *Understanding*, and *Future Actions*. This was repeated 3 times, with different source material in order to trigger diverse questions at different levels of specificity:

1. Participant's own behavior change goals
2. Participant's goals + data screenshots
3. Someone else's goals + data screenshots

3.2 Participants

We held 4 workshops with 12 participants (8 female, 4 male), with an average age of 27.3 (SD=2.9). Three were undergraduate students, 6 graduate students, 3 working as developers and one as a fitness coach. Seven of the participants used Fitbit exclusively, 3 in combination with other tools and 2 used other apps exclusively.

3.3 Reflection Question Generation

The primary goal of the workshops was to generate a set of reflective questions that could be used by a system to trigger reflection. Workshop participants generated a total of 275 questions in 3 categories we prompted for: Noticing (n=76), Understanding (n=116) and Future actions (n=83). Following the generation, we found the questions within one category were not all the same and, in fact, could be further categorized to separate similar and dissimilar question. We decided to perform this categorization to be able to later select the most diverse representatives for each discovered category. To do that, we performed affinity diagramming among 3 researchers to categorize the questions. The most frequent categories are presented in Table 1. These categories represented different specific aspects of behavior change the participant reflected on.

3.4 Insights from Workshop Participants

3.4.1 Current Reflection is Limited and Notice Centric. 7 of 12 participants engaged in reflection on behavior change by mostly reviewing past data and trying to gain self-awareness: *"Comparing / scrolling back" (P7), "Mindfulness about behaviors we don't usually put much thought into" (P6).* Few also reflected to check if they were meeting their goals: *"Am I meeting my goal?" (P4)*, or to gain motivation and a sense of achievement: *"Get additional motivation"(P1), "See my own achievement" (P10).* In terms of frequency, 7 participants indicated they reflect on their data on a daily basis, but mostly on data from the current day only: *"Focus on data at the moment, from current day at most" (P2).* Further, 4 reflected only before or after a specific event: *"After a big workout day, to enjoy achievement" (P7).*

When asked about the ultimate goal of their reflection, 6 considered it to be to provide motivation and to push towards goal achievement: *"Go out to reach daily goal - walk (push towards doing things)" (P2).* Further, 4 considered it to be a self-checking mechanism or a way to gain mental satisfaction: *"Looking at steps helps my self-image when I know I am active" (P6).* Three considered the goal to be learning connections between actions and outcomes and only 1 considered increased awareness to be a goal on its own.

3.4.2 Reflection is Boring, Repetitive and Easy to Forget. 3 of the participants considered reflection to be a boring and repetitive activity: "A little repetitive" (P4), "I get bored with the same goals and stop using the device" (P1). Further, three participants indicated that sometimes they simply forget to look at their data or even to wear their device and would appreciate having engaging triggers that help them reflect: "Having triggers to reflect in an engaging way" (P3), "trigger to think about my data in useful moments" (P9).

Five participants indicated that the metrics provided by fitness trackers do not necessarily help them understand the impact of activities or decide what to do next: *"Need more useful metric than steps/calories" (P3). "Not necessarily clear what to do with it" (P7).* Eight participants suggested the need for better uses of tracked data, including encouraging relevant and insightful comparisons: *"Use comparative analysis for recommendations to set targets" (P2)*, and helping to discover trends and to understand their data better: *"make me understand my data better, specifically what actions work" (P1).* Seven participants indicated interest in receiving suggestions on what to pay attention to, which activities to try, and generally gaining a new perspective on their data: *"getting different perspective, looking at the data in a new way" (P4)*

3.4.3 Self-tracking Data can be Demotivating. A number of participants also reported that they sometimes purposefully avoid looking at their data, and that further reflection can be demotivating: *"I sometimes avoid it because I am worried about what I will see in the data" (P10), "Apps Data motivates or demotivates to achieve goal" (P2).*

3.4.4 Interaction Platform. Mobile phones is the way to go. Regarding the platform of choice for interacting with their self-tracking data, every participant used their mobile phone: *"Access phone/app data 3-4x daily" (P8), "Mobile app for daily checking (everyday at the end)" (P2).* Eight participants used it exclusively: *"Phone only, either at the moment or retroactively" (P9).* Participants would use their mobile phone throughout the day and also look at their data at the end of it. Three participants in addition to their mobile used the display directly on their tracker to do a quick check on their status/progress at the moment: *"Check device in a hurry" (P8), "Watch: real-time checking - see the daily progress" (P10).* Finally, only 1 participant accessed the Fitbit web portal on a laptop/desktop computer: *"Web portal for comparative analysis (once a week)" (P2).*

4 REFLECTION COMPANION: A CONVERSATIONAL SYSTEM FOR REFLECTION

Based on the outcomes of the workshops, we set out to design a system with the following three goals: 1) Guide users towards deeper reflection on physical activity through progressing dialogues, 2) Provide engaging, novel and diverse conversations around reflection, and 3) Allow users to interact with our system on their personal mobile devices.

We designed a system, *Reflection Companion* that engages users through reflection prompts. Reflection Companion uses SMS/MMS for the conversational exchanges. This allowed reaching users regardless of their choice of mobile OS. Reflection Companion initiates a short conversational exchange with an opening question sent once a day at random time within a time range specified by the user. At the enrollment, users

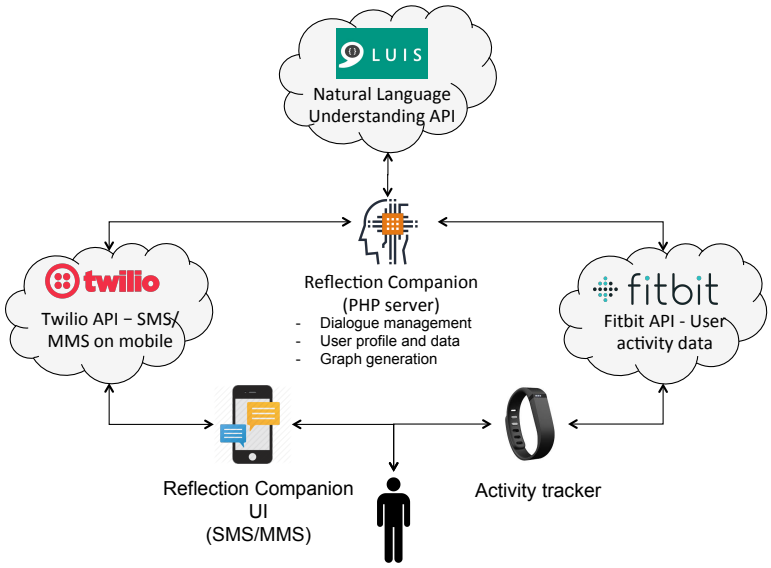

Fig. 3. Overview of the Reflection Companion software components and interaction between them. Twillio API is used to communicate with users mobile phone through SMS/MMS. Fitbit API is queried for user activity data to generate activity graphs. LUIS API offers automated recognition of free-text user responses.

also share their personal daily, weekly and long-term activity goals that are later presented together with a number of mini-dialogues. We implemented our system as a PHP server using a Twilio API for managing the SMS/MMS exchanges (Figure 3). To generate graphs for users' physical activity, we used FitBit API to download latest synchronized user data periodically throughout the day. This required users to grant access to their Fitbit accounts at the beginning of the study (Figure 4). To make the reflection conversation engaging and to encourage a deeper level of reflection, we employed three strategies: the use of a two-step mini-dialogues structure, everyday short reflection sessions, and personalization.

4.1 Guiding Towards Deeper Reflection through Mini-Dialogues

To support deeper reflection, we used a question-follow-up question design, or what we will refer to as a *mini-dialogues* design. Moon suggested the possible use of reflection questions to explicitly guide or structure reflection, further suggesting making use of dialog and discussion [62]. Furthermore, authors in [29] suggested that broad reflection questions could be used to direct thought to the general subject matter, while specific questions could help bring attention to the process of what you are doing in order to learn from it. As an example, authors in [33] prompted students in interactive learning environment to think about what they are doing, provide justifications or explanations for knowledge, actions or events. Based on these indications, our mini-dialogues also have an opportunity to direct the reflection towards deeper levels by bringing users' attention to different aspects of the reflection process. To our knowledge, no technology-based dialogue has been explicitly designed with a purpose of guiding users to a deeper level of reflection on physical activity based on structured reflection models.

To build such mini-dialogues, we created follow-up questions to most of the initial reflection prompts. We followed the progression of the reflection process: questions about awareness would be followed by questions about understanding, whereas questions about understanding would be followed by questions about future actions. In contrast to the initial question where one reminder is sent if the user has not responded within 30 minutes, to minimize interruptions, no reminder is sent after the follow-up question if the user chooses not to respond. 23 out of 25 mini-dialogues feature a follow-up question. The follow-up is asked only after the user provides a response to the initial question. Ten of these have the same follow-up question regardless of what

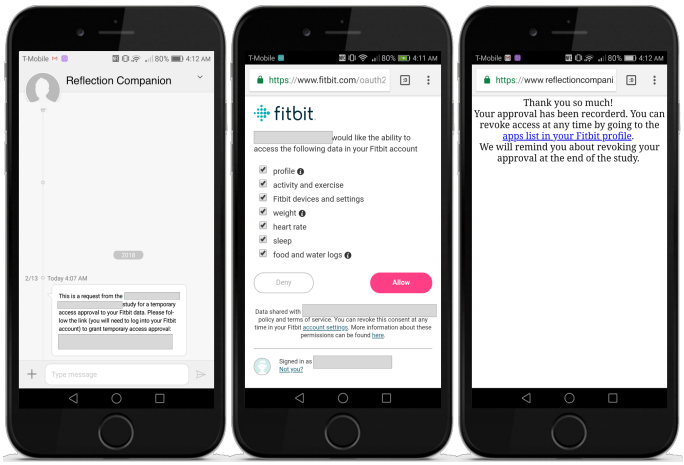

Fig. 4. The interaction on user's mobile phone required for giving the Reflection Companion approval to access Fitbit data.

the user writes in their initial response. However, the remaining 13 mini-dialogues feature a dynamically tailored follow-up question (see dialogues 1 and 3 in Table 2). In such dialogues, a different follow-up question is delivered depending on the user's initial response. The tailored follow-ups are designed in such a way as to build up on user initial response and encourage deeper level of reflection on the shared information, e.g. if initial question asked: *"What are some of the ways that your work has impacted your physical activity this week?"* and user replied with *"Work impacted my exercise because I sit at a desk most of the day"*, then the follow-up question would be *"What could you do to prevent your work from impacting your physical activity?"* On the other hand, if user replied to the same question with: "I walked a lot this week at work because we were changing offices" then the follow-up would be: "How could you set up your work to help you be more active in the future?" In this example the mini-dialogue is trying to guide the user from understanding how the work impacted her activities, to future actions that can help with being more active. This is following the progression suggested by structured reflection models depicted in Figure 2.

Figure 1 shows two examples of the system in action as well as a block diagram of a mini-dialogue structure. As most of the questions from the workshops were formulated with a specific focus on the particular workshop participant they were addressed to (e.g., *"Why do you walk significantly less on Saturdays (and weekends) in general?"*), our first step was to make these questions more general (e.g., *"Why do you walk less on some days?"*). To explicitly support the process of reflection described in past work, we balanced the set of questions focused on noticing, understanding and future actions. To increase the diversity of our dialogues, we selected questions to maximize the number of unique aspects (categories) they belonged to.

As users can provide free-text responses, tailoring the follow-up was based on Natural Language Understanding (NLU) model we trained using the online Language Understanding Intelligent Service (LUIS) [81][1]. Microsoft's online LUIS platform is accessible through REST calls and therefore can be easily linked to any application. We used intent-models to convert the free-text in user response into a machine-readable meaning representation [77,79]. In personal assistant dialog systems, intent-models are classifiers that identify the category of user reply, e.g. *"add meeting to calendar"* or *"play a song"* [82]. Intent models as supervised classifiers require examples of user utterances to be trained on [82]. In the process of training appropriate intent models for our mini-dialogues, we first examined the opening question for a particular mini-dialogue, e.g. *"What are some of the ways that your work has impacted your physical activity this week?"* Given such question, we brainstormed possible classes of responses (intents) users could provide. In this case we decided

[1] https://www.luis.ai/

Table 2. Three sample mini-dialogues used in our system. Reflection-level progression describes the matching of the initial and follow-up questions to the intended reflection levels. For the User-Response intent, the free-text user response is categories using NLU techniques.

#	Reflection-level progression	Mini-Dialogue features	Initial question	User response intent	Follow-up question
1	Noticing → Understanding	Data Graph Goal reference Tailored follow-up	*Referring to your goal of [goal]. How many days did you meet that goal?*	`Goal met` `Goal not met` `Other`	*What did you do on days when you met that daily goal?* *Why didn't you meet your goal on any day?* *Why did you set such a goal for yourself?*
2	Understanding → Future Actions	Static follow-up	*Why is physical activity important for you?*	`Any`	*What steps can you take to be more active?*
3	Future Actions → Understanding	Data Graph Tailored follow-up	*Take a look at your graph. Do you think you can be more active?*	`Yes I can` `No I can't` `Other`	*What small changes (daily repeatable) can you make to be more active?* *What barriers prevent you from being more active?* *Do you want to be more physically active?*

that possible intents are: *"Positive impact"*, *"Negative impact"*, and *"No impact"*. Naturally, the exact way users can express the negative impact of work on physical activity can be vastly different and diverse examples need to be provided for properly training the intent classifier [81]. For that reason, we created various reply examples, such as: *"I sit at the desk or in meetings so that impacts movement"*, *"I had a deadline and I couldn't go to the gym"*, *"It had bad impact"*, *"For example on Tuesday I had very few steps, because I stayed in the office till late"*.

In order to better cover the diversity of possible real-life user responses, we also interacted with the dialogues on a daily basis for around 2-3 weeks among 4 researchers involved in the project. This training and fine-tuning during the development of the Reflection Companion, sometimes also resulted in addition of new intents. For example, for the described mini-dialogue, we did not initially think that "Positive impact" could be an intent and we added it as a result of internal testing. For this particular dialogue we generated 61 different examples of "Negative impact" responses, 58 examples of "No impact" and 38 examples of "Positive impact" responses. In general we created between 30-80 example responses for each intent to be recognized in each mini-dialogue. LUIS provides an interactive learning environment, where the new examples can be entered and used for retraining the existing intent models. In the training phase, 100% of the provided examples were recognized correctly for each mini-dialogue.

There is, however, always a possibility that users provide a response that we did not anticipate. Specifically a response that represents completely different intent than the ones we designed our system for. For that reason, we created an explicit "Other" intent for each dialogue that captured all user responses that didn't

match any other intent (see Figure 1 and Table 2). To train such intent we used a set 100+ various random sentences taken from Wikipedia, Google search results, and random articles, or were self-generated.

4.2 Everyday Reflection Session

An important aspect in the design of our system was the frequency of prompting users to engage in reflective conversations. Too frequent requests for reflection can potentially make the topics to reflect on repetitive and can lead to boredom or frustration given similar activity data and finite diversity of our mini-dialogues [46]. On the other hand, too infrequent reflection can cause people to forget previous revelations, preventing them from building up on past observations and disrupting support for reflection as a process [73].

Human-provided counseling sessions happen infrequently, no more than once or twice a week [74], similar to the frequency of meetings observed in programs such as Weight Watchers [40]. These sessions are, however, much deeper and more extensive than what our Reflection Companion can currently support. Our mini-dialogues are designed to provide brief moments of reflection, rather than support full motivational interviewing sessions. Given indications from past work that users of mobile activity trackers frequently engage in short awareness interaction sessions with their data within one day [35], along further feedback from our workshops, where active tracker users indicated checking their data on their mobile phone at least once a day, we decided to prompt users daily.

4.3 Providing Personally Relevant and Diverse Conversations around Reflection

To make the reflection dialogues engaging, we personalized the experience by introducing questions that *referenced users' own behavior change goals* using an introductory phrase such as: "*Hi Jake, you listed as one of your goals: 'taking regular breaks daily'… "* after which a reflection question would be presented (see #1 in Table 2). The introductory phrases changed each time to provide for a more natural experience. 5 mini-dialogues referenced users' behavior change goals. These mini-dialogues were template based and automatically used the user reported daily, weekly or long-term goal. Each dialogue also addressed users by name and employed a friendly conversational tone following indications from [46].

Furthermore, in order to make the reflection focused and personally relevant, 17 mini-dialogues were delivered with a graph showing user's physical activity metrics (15 plotting steps, one calories burned and one sleep). 14 of these graphs showed a week worth of data, 3 showed a comparison of two weeks of steps (see Figure 2). The ones used in the core 2 weeks of study showed only steps data (see Figure 1). To provide an explicit link between the data shown in the graph and the reflection questions, such mini-dialogues would open with phrase such as: *"Hi Kate, please take a look at your graph…"*. Such introductory phrases again varied each time to provide a more natural experience.

Finally, to diversify the dialogues and to keep users engaged for longer and avoid boredom following indications from [46], we made the dialogues different in terms of the behavior change aspect they addressed. Following the categorization from our workshop presented in Table 1, 8 dialogues were related to observations/patterns, 6 to goals, 4 to plans/schedule, 3 to tracking and general/context, and 1 to motivations. We also diversified them in terms of the starting reflection level - 11 started with noticing, 8 with understanding, and 6 with future actions - and question format - 15 were closed questions and 10 were open questions. This is on top of delivering some of the mini-dialogues with associated activity graphs.

5 FIELD DEPLOYMENT

To evaluate *Reflection Companion's* performance, conversational design choices, and the ability to trigger reflection and encourage participation, we conducted a 2-week field study approved by our university's Institutional Review Board.

5.1 Method

At the start of the study, participants provided our system with access to their Fitbit data. Participants completed a survey, in which they shared their daily, weekly, and long-term behavior change goals and

indicated the time frame during which they would like to receive the reflection mini-dialogues. They then completed a set of scales related to awareness, mindfulness and reflection (detailed in the Measures section), followed by demographics. During the study, participants received one mini-dialogue per day over the course of 2 weeks, delivered to their mobile phones via SMS/MMS. At the end of the 2 weeks, participants completed a post-study survey, responding again to the same scales. Participants also indicated willingness to take part in a phone interview. Finally, participants were allowed to choose to use the system for 2 more weeks without additional compensation (we clarified that their decision would not affect their payment).

5.2 Measures

First, to assess the performance of NLU intent-classifiers used in our system to recognize free-text user responses and match appropriate follow-up questions we looked at the accuracy measure [75]. Furthermore, to assess the success of our design strategy of dealing with user responses that can't be automatically recognized and prevent conversation breakdowns, we qualitatively coded the quality of the dialogue exchanges (Cronbach's alpha, α=0.82 for two independent coders).

Furthermore, to assess the impact and success of Reflection Companion, we looked at measures of engagement. We looked especially at participants' willingness to use the system for additional 2 weeks without compensation as a practical measure of the success of the design of Reflection Companion. Prior work indicates that continuous engagement intention is strongly related to perceived value and satisfaction with the system [43]. Other measures of engagement involved analysis of participant's responses to mini-dialogues and any changes in the self-reported scales. We then examine participants' attitudes and descriptions that emerged from the interviews.

Participant interactions with the system were logged and analyzed. This includes the number of dialogues responded to, the time until a response was made (and whether a reminder was used), as well as the length and content of responses. These measures along with continued participation were used to assess engagement with the system. Participants' daily steps were recorded from their Fitbit data, as well as steps they took the week before the study. We also asked participants to complete a set of scales before and after the study. These included a scale of health awareness adapted for physical activity using a 9-item questionnaire from [39], level of reflection around self-tracking using an adaptation of Kember's 12-item questionnaire with 4 constructs of habitual action, understanding, reflection, and critical reflection, from [42] and general mindfulness using a 13-item questionnaire from [78]. Changes in pre- and post- scale ratings were analyzed using paired t-tests.

To gain a deeper understanding of participants' interaction with the system, user replies to mini-dialogues over two weeks were analyzed and categorized. On top of that, the semi-structured interviews were conducted following the study. Each interview lasted 40 minutes on average and was audio-recorded. The interviews explored the following aspects: 1) general experience with a system, 2) things learned about behavior and tracking, 3) experience of reflection, 4) perception of mini-dialogues, 5) value of every-day questions, 6) feeling of engagement, 7) impact on behavior change goals, 8) impact on motivation/self-efficacy, and 9) reasons for continuing/not-continuing for 2 additional weeks. Interviews were first transcribed and quotes related to each of the categories covered in the interview were extracted following a closed, selective coding approach. Quotes were then regrouped by iteratively subdividing or combining across the initial categories to reveal a set of stable underlying themes. This process required several iterations and followed a general procedure for analysis of qualitative data described in [52].

5.3 Participants

Participants were recruited through social media (17 fitness related Facebook groups, 30 fitness Twitter tags and Reddit) as well as mailing lists and physical flyers. Participants were U.S. based, used Fitbit for at least 2 week, willing to provide access to their Fitbit data, and willing to receive up to 4 SMS/MMS messages per day on their mobile phone for a period of 2 weeks.

Participants were rewarded $30 for the study and an additional $20 for the interview. A total of 33 active Fitbit users participated in our study (29 female, 4 male) between ages of 21 and 60 (M=36.5, SD=11.2). 55% reported having a college degree or being enrolled in college, and a further 27% indicated having a graduate

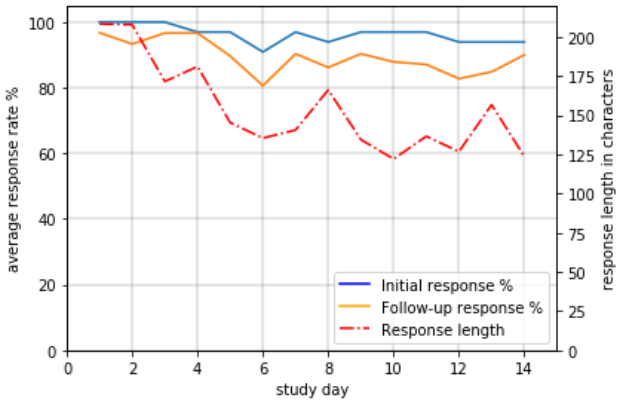

Fig. 5. Response rates to initial and follow-up questions as well as average response length in characters for 14 days of core study.

degree. Participants logged 10,133 steps per day on average (SD=6,521, range: 1768 - 36757) during the week before the study. Five participants logged fewer than 5k and 13 more than 10k steps per day. Furthermore, our participants rated themselves highly on scale of awareness (M=5.63, SD=0.93). While this is a positive characteristic of our participants, it also implies that finding increases in this scale is unlikely - a known outcome or limitation of bounded scales (numerically, participants who rate themselves 7 have no room to improve). 19 of the 33 participants were interviewed after the study.

6 RESULTS

6.1 Engagement with the Reflection Companion

We examined how participants used the system, focusing on the two weeks completed by all participants. During this time, our system sent a total of 462 prompts and 429 follow-ups, receiving 829 responses from participants. Participants responded to 96% of all initial questions and to 90% of the follow-up questions. While 11 participants responded to all questions, the lowest rate for participant responses to initial and follow-up were 23% and 64%, respectively. Overall response rate stayed fairly consistent, indicating generally high engagement throughout the study. However, Figure 5 shows a decline in the length of response as the study progressed, decreasing from an average of 170.1 characters in the first week (SD=31.8) to 138.1 characters in the second week (SD=17.0). Participants took 50 minutes on average to respond to the first question and 13 minutes to respond to the follow-up. Reminders were sent in 39% of cases.

One highly encouraging indicator of our system's viability for ongoing use is the large number of participants who elected to continue to use the system beyond the study's 2-week period. In fact, 16 out of the 33 participants elected to continue using the system for 2 additional weeks without reward. Furthermore, these participants continued to engage with the system at a high rate, responding to 83% of the initial questions and 76% of the follow-up questions during the additional 2 weeks. Average response length during the additional 2 weeks was 98.4 characters (SD=74.9). While the system serves, in part, as a reminder, the sustained high engagement suggests that participants found additional value in the system's use for triggering reflection. We now turn to participants' own description of their experiences using the system.

6.2 System Performance

Our system relied on the underlying NLU classification models to categorize free-text user responses into specific intents (categorizes of user replies) for a subset of mini-dialogues as described in our system design

section. For the 224 replies logged for these dialogues in the core two weeks of the study, more than 72% have been automatically matched with a known intent and resulted in presentation of a tailored follow-up.

We further qualitatively coded the quality of the follow-up question into: *Good match* - follow-up question provided a good continuation of the dialogue, *Acceptable match* – follow-up question only partially build-up on user question or required users to repeat some of the initial response, and *Poor match* – follow-up question made no sense in the context of user reply We found that, when the system was able to automatically recognize user intent, 95% of the presented follow-up questions would be a good (69%) or acceptable match (23%). This means that the system made very few "hard" mistakes, such as recognizing that the user expressed a negative impact of work on physical activity, where in fact user described a positive impact.

For the 62 (22.68%) cases where the system was not able to recognize any intent from user response and for which a non-tailored follow-up was presented, 92% of the presented follow-ups offered a good (58%) or acceptable match (34%), with only 8% of "hard" mistakes.

Not all the dialogues exhibited the same challenge for automated classification. The mean accuracy (a ratio of correctly classified responses to the total number of responses) was 0.69 (SD=0.17). We found that the 2 dialogues with lowest level of recognition accuracy were the ones opening with *"Do your friends exercise more than you do?"* with accuracy of 42%, followed by dialogue opening with *"What are some of the ways that your work has impacted your physical activity this week?"* with accuracy of 53%. Examining the responses revealed that the main reason for misclassification were often ambiguous replies, e.g. for the first question: *"Some do and some don't. I'd guess only 30% exercise more than me."* or for the second question: *"I definitely get more steps on days I commute on public transit. The other days I work from home."* These issues could possibly be mitigated by explicitly designing an intent that models such mixed responses, or by providing additional measure of ambiguity or conflicting information in the user response.

6.3 Impact on Participant's Reflection

6.3.1 Analysis of Reflection in Responses to Mini-Dialogues. In our analysis of user responses to the reflective mini-dialogues, we found that the dialogues were in fact successful in supporting discussions around awareness related to goal accomplishment, self-tracking data, and trends in behavior: *"I like to be active on the weekend and it catches up to me on Mondays so I take it easy, then it's back to working out on Tuesdays and Wed."* Mini-dialogues also appear to have helped participants to better understand their behaviors. They were able to draw connections between the step count and their context, such as weather and external events: *"The weather helped! Also circumstances -- I had meetings and events that I needed to walk to."* And relate physical activity to mental states and lifestyle routines: *"I've been really stressed at work lately, which has made me less active, since I need to finish projects."*

Additionally, participants reflected on multiple higher-level aspects such as the value of physical activity, the meaning of a healthy lifestyle and the value of comparing oneself to others: "*My best friend is a doctor and has 3 kids and exercises way more than I do. (...) So sometimes I feel lazy when I compare myself to a friend, but most of the time I realize this is my life and comparing myself to someone else is not a mentally healthy practice, so I give myself grace."* They often reflected upon things that worked for them: *"Jogging helps me towards the goal of jogging a half marathon. Writing out my training plan on a calendar has been helpful."* and also about what they could possibly change: *"Short runs before or after work. I enjoy running but I don't often make the time anymore. Standing at my desk more. Taking breaks not just at lunch. Getting a dog."*

Aside from reflection, the dialogues provided additional benefits. For example, the prompts enabled users to vent: "Annoyed that some of them are thin without even putting in that much effort. Sometimes annoyed that I can try so hard for less rewards". The mini-dialogues often served as reminders: "Today is my first day back at work so I have not done it yet - will do it if I go to a diff floor".

6.3.2 Pre and Post Quantitative Measures. Looking at the self-reported ratings in Table 3, we find a significant difference in Habitual Action (HA) for pre (M=3.16, SD=1.06) to post (M=3.53, SD=0.89) study measurements; t_{32}=-2.0386, $p<0.05$. We also find a weakly significant increase in Understanding (U) from pre (M=3.60, SD=0.98) to post (M=3.92, SD=0.84); t_{32}=-1.8994, $p=0.07$. Other differences were not significant.

Table 3. Summary of pre- and post study measures. The levels from Kember's survey are mapped to the stages of reflection in the structured reflection process.

Mapping to the stages of a structured reflection process	Measures adapted from Kember [42]	Pre study	Post study
Stage 1: Noticing	Level 1: Habitual action (HA)*	3.16	3.53
Stage 2: Understanding	Level 2: Understanding (U)†	3.60	3.92
Stage 3: Future actions	Level 3: Reflection (R)	3.54	3.64
	Level 4: Critical reflection (CR)	3.60	3.85
Other measures		Pre study	Post study
	Mindfulness	2.52	2.63
	Physical activity awareness	5.63	5.73
	Step count (weekly mean)[1]	10,133	11,165

*Significance against the pre-study measure: * $p<0.05$,* † $p<0.1$,
1 -steps for a week before and after the study as our participants allowed us to stay connected to their Fitbit for an additional time.

The increase in Understanding level indicates an increase in users' analysis of the situation from different perspectives, formulating explanations and observations about the reasons for the things noticed. This result supports our interview findings and is a promising indicator of our overall approach given our study lasting only two weeks. On the other hand, the increase in HA, is a bit harder to interpret. HA is defined as an activity learned in the past that through frequent use becomes something performed habitually [26]. Some prior work has suggested that an increase in HA represents the absence of reflection. We find this interpretation to be unlikely, given that, when reading through participants' reflection responses, we clearly see responses that demonstrate higher levels of reflection.

One likely explanation for the increase in HA is that our system enabled a decoupling of the activity (here, physical activity) from reflecting on the activity (here, taking place when engaging with the system). In the wording of the questions, reflection co-occurred with the activity ("When I am working on some activities, I can do them without thinking about what I am doing"). But since Reflection Companion did not interact with users while they are performing physical activities, it is likely that the decoupling may have occurred and participants responded higher to these HA questions when they used our system.

6.4 Analysis of the Interviews

6.4.1 Types of Reflection Triggered. The 19 interviews confirmed and expanded the results of the analysis of user responses to the mini-dialogues in showing that the system was successful in triggering reflection on past activity patterns, on possible future actions and on new, previously not considered aspects.

Increased Awareness: 10 of the interviewees reported that the system increased their awareness of past physical activity. It specifically helped them realize how much they were recently doing and notice repeatable patterns in their own physical activity: *"It made me more aware that I am doing more steps when I'm at home and on the weekends. It just made me much more aware of how little and how much I'm doing on certain days." (P8).* 4 interviewees claimed the system helped them think about how they currently plan and allocate time to their activities: *"Got me to go back through my data and my calendar, and really stop and spend time thinking about, 'Okay, am I really prioritizing this or not?'" (P14*). Also, 4 participants said it led to them thinking about the relationship between activities, data, and the health outcomes: *"It opened my eyes to a few things… how my steps were affected by what sleep I had…and tracking my patterns on what days I did what." (P10)*

Alternatives and Future Actions: 8 interviewees said that interacting with the system led to reflection on the actions they were currently taking to achieve their goals and made them critically re-evaluate these actions to think about possible alternatives: *"I definitely thought about whether I was doing as much as I could to be able to*

reach those goals. More about what were the barriers that were making it where I wasn't reaching those goals." (P13). The prompts also triggered thinking about planning possible strategies to achieve enough physical activity based on what they have learned from the past: *"Partially, it's about reflection, but it's more of planning ahead, like what I should do and what I will do…by reflecting on the past behavior." (P20).* Such reflection was for many participants a prerequisite for trying out new behaviors.

New Insights: 4 participants said interacting with the system led to reflection on aspects they had not thought of before, such as considering possible alternative metrics: *"It got me thinking about what other interesting metrics are there? I had never really thought about what I track or pay attention to that carefully. I just kind of use whatever the given dashboard is." (P14).* In other cases, it triggered critical thinking about how they currently use the metrics that are tracked, and what they can learn from these metrics. The system also introduced new ways to evaluate data by presenting them in a different timeframe (e.g., two weeks): *"It was my first time to see an overview of my weekly activity…I had never done it before. Thinking in a way of a week cycle was interesting...Thinking of two weeks in parallel, is there any seasonality or any cycle." (P20).*

6.4.2 Benefits of Reflection. We found that reflection was beneficial in many ways: by increasing motivation towards physical activity, introducing changes to participants' actual behavior, increasing mindfulness, and encouraging the formulation of more realistic strategies for increasing physical activity.

Increased Motivation: Many participants found the reflective dialogues to be motivating. 5 interviewees reported that the mere presence of the prompting mechanism provided focus, kept them in check, and consequently led to increased motivation: *"They pushed me, in my opinion, they pushed me to start doing more. And sometimes we need that little push." (P3).* In some cases, the daily presence of the dialogues created a sense of accountability, which provided additional motivation: *"They were a form of encouragement to me, because it's like I knew that there was accountability on my part, that if I had a poor day that I had to explain why, reflect on that on, what would I do the next day." (P22).* 8 interviewees reported that the dialogues helped them realize their barriers, formulate clear action plans and define small, concrete and attainable steps for achieving their goals. Interviewees considered these aspects to be motivating: *"It was like 'What little changes could I do?' And that was helpful 'cause like making the time for an hour workout every day seems daunting, but going for a walk on my lunch is doable. Going for a walk after work is doable." (P25).*

Leading to New Behaviors: For many interviewees, engaging in reflection resulted in the adoption of new behaviors. These behaviors were usually small changes to daily routines, such as parking further away from office or parking meter to walk more, walking to a grocery store instead of taking a car, or using stairs instead of an elevator: *"I actually did little things to make myself more active during the day. The prompts got me like, one day I'm talking about walking more during break, and so since then I've made a point to get out of the office and walk during my lunch. Just doing little things." (P25).* In some cases, the dialogues served as an additional push on top of a request from a family member, e.g. a request from participant's daughter to go for a walk or an evening walk with wife in case of another participants. In some cases, the prompts also triggered a return to past behaviors that have been abandoned: *"It actually got me to get back into running, which is what I had gotten out of for a little while so that was kind of nice." (P24).* In a number of cases, the mini-dialogues led to behaviors that facilitate physical activity, such as wearing Fitbit more often, downloading an additional app for tracking running progress or scheduling a class at the gym: *"After I would get the message, if I hadn't already scheduled class at the gym for that day, it would usually be a good reminder." (P14)*

Increased Mindfulness and Leading to More Realistic Plans: 6 of the interviewees said that the mini-dialogues helped them better assess their progress and become more mindful of their own tendencies and inclinations: *"I realized something about myself that I like to work out…[by doing] another activity. For example, going to the museum." (P14).* In many cases, this led to an increased understanding of factors that help participants meet their goals, or barriers that prevent participants from reaching their goals: *"I guess just becoming more aware of the barriers to some of the stuff keeping me from my goals." (P26).* This helped interviewees realize the need for specific and realistic actions to achieve their goals: *"I think it helped me be more realistic. A lot of times where*

Table 4. Summary of the positive/negative aspects of the system design choices based on feedback from participants.

	Aspects of mini-dialogue based reflection guidance		**Reflection frequency: One dialogue a day**	**Aspects of Personalization & diversification**
	Two-step mini-dialogue structure	**Typing and sending responses**		
Positive aspects	Extends thinking time for reflection	Promotes deeper thinking, seriousness, and precision	One dialogue a day just right: allows for reflecting on continual progress	Graph useful for supporting response closely tied to the personal data
	Encourages deeper thinking and more meaningful answers	Creates a sense of commitment and accountability	Enables devoting the whole day for reflecting on one aspect, which was appreciated	Prompts useful for bringing attention to the data aspects not considered before
	Having two smaller questions lowered the reflection effort	Helps remembering, serves as a mental note	Useful as a momentary trigger and check-in	Personal data promoted engagement and motivation
Observations /Challenges	Some follow-up questions felt generic and computerized	Required additional effort from the user	Aspects discussed between dialogues are sometimes repeated	Despite diversification graphs and questions felt similar

you're like 'Oh I can do this in a month or something like that.' But in reality, it's a lot tougher so it's nice to have that reflection" (P24).

6.4.3 Impact of System Features. In this section we focus on exploring the impact of key elements of the system: the two-step mini-dialogue structure, continuous reflection through daily conversations, the need for typing and sending a response, and personalization using the activity graph of personal Fitbit data (Table 4).

Two-step Mini-Dialogue Format: Our dialogues were composed of an initial question and a follow-up question, which were overall positively received by interviewees. They felt that the follow-up questions gave them a chance to spend more time reflecting on the initial question, noting that one question was not enough to engage in meaningful reflection: *"t think if just one prompt might be too short for me to reflect on my activity. I think I need at least like a minute or so experience to really think about how I feel and why I did it and things like that." (P20).* Alternatively, some participants considered the initial question to be a warm-up to the follow-up question, which encouraged them to answer more truthfully: *"I mean I think it caused me to answer more truthfully, more honestly, or more alertly, so I couldn't just give a one-word answer or anything, I have to kind of think about it a little bit more." (P28).*

Future work is still needed to help make the interaction more dynamic. While the follow-up questions were generally well-received, a few interviewees did not consider the follow-ups to be sufficiently adaptive, which caused the mini-dialogues to feel computerized: "*I would write the response, and then get a generic response back. It was like, "Are you serious?" It felt a waste of my time, writing a long text." (P8).*

Daily Prompting: All interviewees appreciated receiving one prompt per day. Specifically, 6 felt that having a daily reflective dialogue enabled them to view and reflect on their continual progress: *"Everyday, I am able to see the progress. I am able to reflect on the previous day or the previous week." (P3).* Four participants also

expressed that it helped them focus on a specific day, which lowered the cognitive load of reflection. Furthermore, some participants reported that reflection persisted beyond the dialogue: *"I almost got a whole day thinking about one question, even after I'd already sent the responses out. Which then allowed me to build upon what I was thinking about 'cause of repetition." (P25).* Finally, daily delivery served as a momentary trigger (e.g., to put Fitbit on or to remind them to walk more). When asked about the frequency of prompts, one participant said: *"You don't want it to go too long in between, because then it starts to feel random. But if you have something that's checking in on you once a day, then it's a way to just check in. It becomes part of the routine." (P32).*

Typing and Sending Responses: 7 interviewees stated that writing the response to the reflective questions felt like an additional reinforcement on top of simply thinking about the answer. It caused them to think deeper and forced them to put their thoughts into words. This can act as a commitment device and create a sense of accountability to self: *"It gives you a little sense of accountability…When you take the effort to actually put what you did down, you kind of reflect on your performance. You know if you're lying to yourself or not." (P10).* For 6 interviewees, the sense that someone (computer or person) was reading their responses led to being more conscious about what they wrote and to a sense of accountability. This happened despite the participants acknowledging that the 'someone' is a computer program: *"It made me feel like I was being accountable to somebody, even if it was just a computer program. So I liked that. It made me more motivated." (P23*). Finally, as a record of exchanges was kept on participants phone, typed response also served as a reminder or a note: *"it's like when I write something down, you're able to remember it better. Because I was consistently typing it on my phone or making a mental note of it, it was a good reminder." (P24)*

Reflection with Data: 11 interviewees expressed appreciation for the graphs included with the reflection prompts and considered them crucial for awareness and revealing progress: *"I think them being conjoined was helpful. The graph and then asking me a specific question about that data. Then I really had something to tie my answer to" (P27).* Aside from helping users tie their response to the graph, the reflection dialogues helped focus user attention on aspects of the graphs they have never considered before, even despite having access to the same graphs in the Fitbit app: *"Like some of the graphs that you guys have sent me… I hadn't really thought about the patterns and so that has been useful for me to actually think about that more and I should probably have different step goals for weekdays than I do for weekends." (P23).* In some cases the visual and personalized nature of the graph was considered particularly engaging and provided additional motivation. Some participants, however, reported the graphs to be redundant given that their Fitbit app already provides a similar visualization.

7 LIMITATIONS

One limitation of our 2-week deployment is that it may have been too short to allow users the transition through the full stages of the reflection process. Indeed, reflecting and eventually integrating the results of the reflection into behavior can take a long time. Still, we were able to show that even during these two weeks, all users' reflection metrics have increased, albeit only an increase in habitual action and understanding - an early stage of the reflection process - reached significance.

Another limitation is that some of the effects we observed might be attributed to novelty of the system. While we can't rule that out completely, with half of the participants continuing to use and engage with the system beyond the 2-weeks, *Reflection Companion's* potential for longer-term use seems likely.

The final limitation is the lack of a clear control group, as our baseline was the active use of Fitbit by all the participants at least 2 weeks prior to the study. This opens up a possibility that perhaps a simpler setup, e.g. just reminders, could have achieved similar effects to the ones we observed. However, the qualitative results, with participants being able to point to specific conversational design aspects that were beneficial, gives us confidence this is not the case.

8 DISCUSSION

Quantified-self technologies can collect massive amounts of data about the user, yet it has been shown that just having access to data is not tantamount to learning valuable information from it [27,38]. In this respect, reflection has been identified as one of the key processes that can support such learning [27], which can be applied in behavior change context [56,59], as well in other domains [6,44]. Unfortunately, most existing tools assume that reflection would naturally occur when people visualize data or journal, which has shown not to be the case [29]. In our work, we argue that a conversational approach, using what we refer to as "mini-dialogues" design, can be effective for eliciting reflection. Indeed, in our deployment, our system successfully led to reflection at three levels: awareness, understanding, and new insights for the future. We show that such reflection can help users become more motivated and can lead to defining action plans better aligned with users' long-term goals and actual abilities. Here we further discuss some of the aspects of our approach. We also summarize the lessons learned and key design insights into Table 5.

8.1 Benefits and Drawbacks of Reflection on Physical Activity

In accordance with prior work [17], our workshops with active Fitbit users identified that reflection these users engage in is often limited, with a current lack of proper tools support such reflection. Many reported that, when reflecting, they focus predominantly on a day's worth of data or reflect on a single event. They reported a number of barriers that limit their engagement, from perceiving that reflection can be boring and repetitive, to lacking engaging triggers to reflect on novel aspects of their data and activity. With our system, we demonstrate that dynamic, daily triggers of reflection may be able to fill this gap. Here we discuss some of the benefits and drawback of reflection in the context of physical activity we identified in our study.

One of the main benefits is that reflection *helps increase awareness, mindfulness, and consideration of new aspects.* Much prior work has discussed the importance of reflection in behavior change [54], has suggested that reflection can lead to greater awareness of underlying needs [53], has helped people understand their motivations and focus on higher-level goals [14,76]. It can also help overcome decision biases and give opportunities to better focus on the actions [61]. Reflection can further help people focus on long-term consequences of their choice and make decisions that are more in line with their identities [3], which also makes them more likely to tolerate short term discomfort [32]. In our work, this is supported both through our interview data, as well as the increase in Understanding ratings pre and post study. On top of these benefits we also found that diverse prompts can help people consider new perspectives and interpretations of their personal activity data they thought they are already familiar with. Reflective prompts can serve as a powerful tool for exploration.

We also found that reflection activities *can serve as a prerequisite to better goal setting and more feasible future actions.* In alignment with prior work, indicating that reflection on physical activity goals prior to setting such goals can lead to eventual increases in physical activity [53], we found that reflection serves as a preparation for considering new goals and feasible future actions. At the same time, we confirmed indications from prior work, that reflection is a slow process that requires time [37]. While most of our participants did not revise their physical-activity goals during the study, many reported that the 2-week period was too short to compel such a revision.

Further, we found that reflection *provides a non-judgmental, neutral interaction.* On one hand, the reflection activities offer participants a break from the often judgmental and persuasive nudges built into behavior change systems. This was especially appreciated by a number of participants concerned with their current level of physical activity. However, for some others, a concern was that reflection activities are *not necessarily actionable.* In the context of physical activity, many users expected more motivational and actionable support. Reflection helped participants think about possible actions, but as P21 puts it "*I didn't actually do it, but I thought harder about it.*"

Finally, we should note that reflection *might potentially lead to discouraging revelations.* During our workshop sessions we found indications that looking at the data may lead to discouraging observations (e.g., noticing that one is not as active and feeling bad about it). We designed our dialogues to help guide users to move on to next stage of reflection to avoid being stuck in negative thinking. Encouragingly, we did not

notice any indications of our prompts having such negative effects during the study, but this still stays a remote possibility.

8.2 Insights About Designing a Conversational Agent for Reflection

Through our study, we uncovered three key benefits of the conversational approach. One is that it *has an ability to actively shape the direction of user thinking*. We found that the dialogues, through subsequent prompts building-up on user response, have an ability to guide user thinking in a specific direction. In our case, we wanted the users to progress along the stages of reflection process. In that respect, conversational approach is well suited for supporting structured reflection, as advocated in [29]. We also found that having conversational exchanges *extends the time user spends reflecting*. Another benefit is that *everyday conversations can help users learn over time.* Through on going, every-day short conversations, we found that the conversational approach helps users build up on what they have learned before. This is particularly valuable if reflection is approached as an ongoing developmental process [36] for which conversational agent seems particularly well suited. Last but not least, the conversational approach *provides an engagement boost through perceived accountability and commitment*. The act of typing and committing to an answer brings benefits of precision in expression, deeper thinking, and accountability. We found that participants through typing responses would commit to self or to 'someone' reading them, even if they knew this 'someone' was a computer program. This is consistent with the paradigm of computers as social actors we discussed in the related work [64].

However, there are also drawbacks with using the conversational approach. One, doing so r*uns the risk of building-up and disappointing user expectations*. Conversation is a phenomenon naturally associated with intelligence [67]. As automated conversational system is not truly intelligent, the illusion of intelligence is easily broken sooner or later. If the conversational system triggers too high expectations of intelligence for the user, it can easily lead to disappointment and eventual abandonment. Second, conversational interfaces are at least currently *harder to design for; more effort and resources are required*. Crafting conversational interaction, especially with an ability to adapt to user responses, is a much more elaborate design process, than sending disconnected random prompts.

One key challenge with building a conversational system for reflection (or a conversational system in general) is to generate a set of sufficiently diverse and topic-appropriate dialogues. This is especially important for the purpose of continuous, everyday coach like interaction [11,12]. In this work, we used a workshop-based approach, akin to participatory design to address this challenge. In general, we found this to be an effective approach; different users had different data and goals, which helped generate a diverse set of prompts for reflection. Using real data also helped ensure the prompts remain relevant and interesting to potential users. While we believe this to be a valuable approach that can be adopted by designers the matter of further extending the usefulness of the reflection dialogues for long-term use requires a dedicated discussion, which we provide in the next section.

8.3 Extending the Long-Term use of Reflection Companion

Going forward, in order to make the dialogues even more engaging, especially for longer-term use, we discuss a number of potential approaches such as *diversification*, *tailoring*, and *memory & adaptation*.

Diversification focuses on making the dialogues novel each time. It can be applied on *syntactic* (sentence composition), *semantic* (topics), and *dialogue structure* level. Syntactic diversification is perhaps the easiest and we already applied it to some extent (e.g., we used 12 different sentence-openers and 10 follow-up introductions). One could relatively easily introduce new templates and mix them up each time to diversify sentence composition. This is valuable, but limited [71]. A more elaborate approach is further diversification on a semantic level, the level of topics, with additional workshops involving different user groups (e.g. novice trackers) or different set of generation prompts. Also recent work on peoples' insights from reflection on quantified-self data can provide such new topics [18]. Finally, the organization of the dialogue itself can also be diversified, e.g., some dialogues could feature 3 or 4 stages of exchange. Diversification, however, does not build up on past exchanges or knowledge of user interests to make the conversation more engaging.

Table 5. Summary of system design insights grouped by different elements of our conversational system.

	Design for Natural Language Understanding	Design for Everyday Mini-Dialogues	Design for Reflection on Physical Activity
Design insights	**Generate as complete and diverse training data as possible**: diverse training examples and dialogue scoping makes free-text user response recognition feasible in practice.	**Appearing "intelligent" is not always the most important**: despite users quickly realizing they interact with a computer system, they still felt interaction to be valuable and engaging.	**Contextualize dialogues around user personal data**: presenting personal data along with dialogues helps contextualize responses and boost motivation.
	Explicitly handle the unknown: by explicitly training the intent-classifier to recognize "other" intents, we were able to handle such cases with proper design on the dialogue level.	**Design explicit fallback strategy into the dialogues**: providing generic follow-up in failed recognition cases allows avoiding hard conversation breakdowns.	**Generic prompts can still be valuable**: even providing a broad, non-tailored follow-up could lead to deep reflection as users can make their own interpretations.
	Handle mixed-intent responses: most "hard" misrecognitions of user replies were due to the response incorporating aspects of two known intents, e.g. "yes and no". Explicitly designing for such intents can be helpful.	**Consider timing of the exchanges**: chatbots usually respond to users as soon as they process user message, we suggest that it is desirable to delay the exchange to allow users to think, but also handle potential follow-ups users would occasionally send.	**Users are willing to type more in reflection context**: even for closed questions, users are willing to share more, usually starting with e.g. yes/no and then elaborating.
Future design directions	**Identifying new intents on the fly with crowdsourcing**: handling unknown intents requires updates to the dialogue structure, which might be possible using crowdsourcing.	**Consider history of user responses**: take into account what user already shared and present dialogues that are likely to elicit different responses to avoid repetition.	**Transitioning from pure reflection into motivating action**: users expected reflection to also increase their motivation and eventually lead to actions.

Another approach involves personalization & tailoring. We used personalization by addressing user by name, presenting graph of personal data, and weaving in user goals into selected mini-dialogues. The topics introduced by the dialogues were, however, not tailored to user's interests in any way. Tailoring could prompt user to reflect on the topics of interest and inform diversification around such related topics. Schwartz's 10 basic values, which represent universal motivational constructs, could be used [72]. One's values can be assessed with a survey [20] or based on language use [16]. In practice, user oriented towards *achievement*, could receive additional dialogues triggering reflection on achievements, ambitions and capabilities. While user oriented towards *benevolence* might be more interested in reflecting on how activity can be helpful to others or how to enhance one's spiritual life. Value based tailored diversification has been found effective [46]. Yet another option could be tailoring the dialogue structure itself, which has been explored for cultures [24].

Arguably most valuable for long-term, but also most technically challenging, would be to *remember* aspects user shared and adapt the mini-dialogues. Currently user response to the initial prompt is classified and "remembered" to decide on the follow-up to present. Unfortunately, no long-term memory or *common ground [21]* is retained. This requires asking each time e.g. what is user barrier for a specific goal or activity, or having to switch to a new topic to avoid repetition. Remembering such information from user past responses (e.g. a

barrier of *"not having a person to run with"*), would enable reflecting on aspects of this particular and personal information, e.g., *"What could you do to try to find someone to run with?"* or *"Is not having someone to run with still an issue for you?"* This has obvious long-term benefits: it allows to deepen the reflection on relevant topics over time, it communicates to the user that the shared information is appreciated, and it partially addresses the issue of topics exhaustion as dialogues can also go in depth over time. The aspects worth remembering, and bringing back to the user, could be selected based on domain knowledge. Personal data concepts to remember could involve: *external context of activity*, *identified trends*, as well as *outliers* [18], and behavior change concepts could be: *social factors*, *motivators*, *barriers*, *past activities*, and *attitudes* [2]. From a technical perspective, of-the-shelf NLU systems such as LUIS we used, can extract custom trained *entities* from-free text. Finally, the long time between subsequent mini-dialogues also permits the use of crowd-sourcing [23].

9 CONCLUSION

This paper introduced a conversational system based on mini-dialogues for supporting reflection on physical activity. Our workshop results and deployment findings offer many important insights about reflections for personal informatics and behavior change and how conversational interaction can be designed and used to support the reflection process.

In future work, we plan to make our conversational approach more tailored and able to evolve over time to support users' increasing levels of reflective thinking. We are also considering improvements to our dialogue generation process that may streamline and integrate it with the reflection system.

ACKNOWLEDGMENTS

We would like to thank Elena Agapie, Mia Minhyang Suh, and Chia-Fang (Christina) Chung for their help with testing and providing feedback on the early versions of the Reflection Companion. We would also like to thank Jennifer Turns for providing inspiring resources about reflection and sharing valuable insights. This work was in part supported by National Science Foundation grant #1348543.

REFERENCES

1. Jakob Åberg. 2017. *Chatbots As A Mean To Motivate Behavior Change: How To Inspire Pro-Environmental Attitude with Chatbot Interfaces*. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:1106358
2. Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2: 179–211.
3. George A. Akerlof and Rachel Kranton. 2010. Identity economics. *The Economists' Voice* 7, 2. Retrieved from https://www.degruyter.com/downloadpdf/j/ev.2010.7.2/ev.2010.7.2.1762/ev.2010.7.2.1762.xml
4. Sue Atkins and Kathy Murphy. 1993. Reflection: a review of the literature. *Journal of advanced nursing* 18, 8: 1188–1192.
5. John D. Bain, Roy Ballantyne, Jan Packer, and Colleen Mills. 1999. Using journal writing to enhance student teachers' reflectivity during field experience placements. *Teachers and Teaching* 5, 1: 51–73.
6. Scott Bateman, Jaime Teevan, and Ryen W. White. 2012. The search dashboard: how reflection and comparison impact search behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1785–1794.
7. Eric P.S. Baumer. 2015. Reflective Informatics: Conceptual Dimensions for Designing Technologies of Reflection. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 585–594. https://doi.org/10.1145/2702123.2702234
8. Eric PS Baumer, Vera Khovanskaya, Mark Matthews, Lindsay Reynolds, Victoria Schwanda Sosik, and Geri Gay. 2014. Reviewing reflection: on the use of reflection in interactive system design. In *Proceedings of the 2014 conference on Designing interactive systems*, 93–102. Retrieved January 31, 2017 from http://dl.acm.org/citation.cfm?id=2598598
9. Bridgette M. Bewick, Karen Trusler, Michael Barkham, Andrew J. Hill, Jane Cahill, and Brendan Mulhern. 2008. The effectiveness of web-based interventions designed to decrease alcohol consumption—a systematic review. *Preventive medicine* 47, 1: 17–26.
10. Timothy Bickmore and Toni Giorgino. 2006. Health dialog systems for patients and consumers. *Journal of biomedical informatics* 39, 5: 556–571.

11. Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2010. Maintaining Engagement in Long-Term Interventions with Relational Agents. *Applied Artificial Intelligence* 24, 6: 648–666. https://doi.org/10.1080/08839514.2010.492259
12. Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 2: 293–327.
13. David Boud, Rosemary Keogh, and David Walker. 2013. *Reflection: Turning experience into learning*. Routledge. Retrieved from https://books.google.nl/books?hl=en&lr=&id=XuBEAQAAQBAJ&oi=fnd&pg=PP1&dq=Promoting+reflection+in+learning,+in+Reflection:+Turning+experience+into+learning.+&ots=TuWs2Rrg0P&sig=KoFpLEseuZJEdzrg0P4JbEdo_JY
14. Charles S. Carver and Michael F. Scheier. 2000. Autonomy and self-regulation. *Psychological Inquiry* 11, 4: 284–291.
15. Charles S. Carver and Michael F. Scheier. 2001. *On the self-regulation of behavior*. Cambridge University Press.
16. Jilin Chen, Gary Hsieh, Jalal U. Mahmud, and Jeffrey Nichols. 2014. Understanding Individuals' Personal Values from Social Media Word Use. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (CSCW '14), 405–414. https://doi.org/10.1145/2531602.2531608
17. Eun Kyoung Choe, Bongshin Lee, Matthew Kay, Wanda Pratt, and Julie A. Kientz. 2015. SleepTight: low-burden, self-monitoring technology for capturing and reflecting on sleep behaviors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 121–132. Retrieved from http://dl.acm.org/citation.cfm?id=2804266
18. Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding Self-Reflection: How People Reflect on Personal Data through Visual Data Exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth'17). ACM, New York, NY, USA*.
19. Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, 1143–1152.
20. Jan Cieciuch and Shalom H. Schwartz. 2012. The number of distinct basic values and their structure assessed by PVQ–40. *Journal of personality assessment* 94, 3: 321–328.
21. Herbert H. Clark. 1996. *Using language*. Cambridge university press.
22. Sunny Consolvo, Predrag Klasnja, David W. McDonald, Daniel Avrahami, Jon Froehlich, Louis LeGrand, Ryan Libby, Keith Mosher, and James A. Landay. 2008. Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. In *Proceedings of the 10th international conference on Ubiquitous computing*, 54–63. Retrieved from http://dl.acm.org/citation.cfm?id=1409644
23. Justin Cranshaw, Emad Elwany, Todd Newman, Rafal Kocielnik, Bowen Yu, Sandeep Soni, Jaime Teevan, and Andrés Monroy-Hernández. 2017. Calendar. help: Designing a Workflow-Based Scheduling Agent with Humans in the Loop. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2382–2393.
24. Beant Dhillon, Rafal Kocielnik, Ioannis Politis, Marc Swerts, and Dalila Szostak. 2011. Culture and facial expressions: A case study with a speech interface. In *IFIP Conference on Human-Computer Interaction*, 392–404. Retrieved January 10, 2017 from http://link.springer.com/chapter/10.1007/978-3-642-23771-3_29
25. Carlo C. DiClemente, Angela S. Marinilli, Manu Singh, and Lori E. Bellino. 2001. The role of feedback in the process of health behavior change. *American journal of health behavior* 25, 3: 217–227.
26. Mica R. Endsley. 1997. The role of situation awareness in naturalistic decision making. *Naturalistic decision making* 269: 284.
27. 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*, 731–742. Retrieved February 2, 2017 from http://dl.acm.org/citation.cfm?id=2804250
28. John H. Flavell. 1979. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist* 34, 10: 906.
29. 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*, 216–223. Retrieved January 10, 2017 from http://dl.acm.org/citation.cfm?id=1952269
30. Marcus Foth, Jaz Hee-jeong Choi, and Christine Satchell. 2011. Urban informatics. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, 1–8.
31. Michael Friedewald and Oliver Raabe. 2011. Ubiquitous computing: An overview of technology impacts. *Telematics and Informatics* 28, 2: 55–65.
32. Kentaro Fujita and H. Anna Han. 2009. Moving beyond deliberative control of impulses: The effect of construal levels on evaluative associations in self-control conflicts. *Psychological Science* 20, 7: 799–804.

33. Claudia Gama. 2001. Helping students to help themselves: a pilot experiment on the ways of increasing metacognitive awareness in problem solving. *Proc. of New Technologies in Science Education.*
34. Graham Gibbs. 1988. *Learning by doing: A guide to teaching and learning methods*. Oxford Centre for Staff and Learning Development, Oxford Brookes University.
35. Rúben Gouveia, Evangelos Karapanos, and Marc Hassenzahl. 2015. How do we engage with activity trackers?: a longitudinal study of Habito. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 1305–1316.
36. K. Gustafson and W. Bennett. 1999. Issues and difficulties in promoting learner reflection: Results from a three-year study. *WWW: http://it. coe. uga. edu/~ kgustafs/document/promoting. html.*
37. Lars Hallnäs and Johan Redström. 2001. Slow technology–designing for reflection. *Personal and ubiquitous computing* 5, 3: 201–212.
38. Katrin Hänsel, Natalie Wilde, Hamed Haddadi, and Akram Alomainy. 2015. Challenges with current wearable technology in monitoring health data and providing positive behavioural support. In *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, 158–161.
39. Hyehyun Hong. 2009. Scale development for measuring health consciousness: Re-conceptualization. In *12th annual international public relations research conference, Holiday Inn University of Miami Coral Gables, Florida*. Retrieved from http://www.instituteforpr.org/wp-content/uploads/ScaleDvlpmentMeasuring.pdf
40. Kate Jolly, Amanda Lewis, Jane Beach, John Denley, Peymane Adab, Jonathan J. Deeks, Amanda Daley, and Paul Aveyard. 2011. Comparison of range of commercial or primary care led weight reduction programmes with minimal intervention control for weight loss in obesity: Lighten Up randomised controlled trial. *BMJ* 343: d6500. https://doi.org/10.1136/bmj.d6500
41. Daniel Kahneman. 2003. Maps of bounded rationality: Psychology for behavioral economics. *The American economic review* 93, 5: 1449–1475.
42. David Kember, Doris YP Leung, Alice Jones, Alice Yuen Loke, Jan McKay, Kit Sinclair, Harrison Tse, Celia Webb, Frances Kam Yuet Wong, Marian Wong, and others. 2000. Development of a questionnaire to measure the level of reflective thinking. *Assessment & evaluation in higher education* 25, 4: 381–395.
43. Young Hoon Kim, Dan J. Kim, and Kathy Wachter. 2013. A study of mobile user engagement (MoEN): Engagement motivations, perceived value, satisfaction, and continued engagement intention. *Decision Support Systems* 56: 361–370.
44. Rafal Kocielnik, Daniel Avrahami, Jennifer Marlow, Di Lu, and Gary Hsieh. Designing for Workplace Reflection: A Chat and Voice-Based Conversational Agent. *Proceedings of the 2018 Conference on Designing Interactive Systems, DIS '18*: 14. https://doi.org/10.1145/3196709.3196784
45. Rafal Kocielnik and Gary Hsieh. 2018. Facilitating Self-learning in Behavior Change Through Long-term Intelligent Conversational Assistance. In *23rd International Conference on Intelligent User Interfaces*, 683–684.
46. Rafal Kocielnik and Gary Hsieh. Send Me a Different Message: Utilizing Cognitive Space to Create Engaging Message Triggers. *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing.*: 2017.
47. Rafal Kocielnik, Fabrizio Maria Maggi, and Natalia Sidorova. 2013. Enabling self-reflection with LifelogExplorer: Generating simple views from complex data. In *Pervasive Computing Technologies for Healthcare (PervasiveHealth), 2013 7th International Conference on*, 184–191. Retrieved from http://ieeexplore.ieee.org/abstract/document/6563924/
48. Rafal Kocielnik, Mykola Pechenizkiy, and Natalia Sidorova. 2012. Stress analytics in education. In *Educational Data Mining 2012*.
49. Rafal Kocielnik and Natalia Sidorova. 2015. Personalized stress management: enabling stress monitoring with lifelogexplorer. *KI-Künstliche Intelligenz* 29, 2: 115–122.
50. Rafal Kocielnik, Natalia Sidorova, Fabrizio Maria Maggi, Martin Ouwerkerk, and Joyce HDM Westerink. 2013. Smart technologies for long-term stress monitoring at work. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, 53–58.
51. David A. Kolb. 2014. *Experiential learning: Experience as the source of learning and development*. FT press. Retrieved from https://books.google.nl/books?hl=en&lr=&id=jpbeBQAAQBAJ&oi=fnd&pg=PR7&dq=Experiential+learning:+experience+as+the+source+of+learning+and+development&ots=Vn2SnW_XJc&sig=M3cn3_8BZgR77wpUjCmgRyMQwy8
52. Margaret D. LeCompte. 2000. Analyzing qualitative data. *Theory into practice* 39, 3: 146–154.
53. Min Kyung Lee, Junsung Kim, Jodi Forlizzi, and Sara Kiesler. 2015. Personalization revisited: a reflective approach helps people better personalize health services and motivates them to increase physical activity. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 743–754. Retrieved from http://dl.acm.org/citation.cfm?id=2807552

54. 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*, 557–566. Retrieved from http://dl.acm.org/citation.cfm?id=1753409
55. 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*, 405–414. Retrieved January 10, 2017 from http://dl.acm.org/citation.cfm?id=2030166
56. Ian Li, Jodi Forlizzi, and Anind Dey. 2010. Know thyself: monitoring and reflecting on facets of one's life. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, 4489–4492. Retrieved from http://dl.acm.org/citation.cfm?id=1754181
57. James Lin, Lena Mamykina, Silvia Lindtner, Gregory Delajoux, and Henry Strub. 2006. Fish'n'Steps: Encouraging physical activity with an interactive computer game. *UbiComp 2006: Ubiquitous Computing*: 261–278.
58. Madelene Lindström, Anna St\a ahl, Kristina Höök, Petra Sundström, Jarmo Laaksolathi, Marco Combetto, Alex Taylor, and Roberto Bresin. 2006. Affective diary: designing for bodily expressiveness and self-reflection. In *CHI'06 extended abstracts on Human factors in computing systems*, 1037–1042. Retrieved from http://dl.acm.org/citation.cfm?id=1125649
59. Lena Mamykina, Elizabeth Mynatt, Patricia Davidson, and Daniel Greenblatt. 2008. MAHI: investigation of social scaffolding for reflective thinking in diabetes management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 477–486. Retrieved from http://dl.acm.org/citation.cfm?id=1357131
60. Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. 2012. AffectAura: an intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 849–858. Retrieved from http://dl.acm.org/citation.cfm?id=2208525
61. Katherine L. Milkman, Dolly Chugh, and Max H. Bazerman. 2009. How can decision making be improved? *Perspectives on psychological science* 4, 4: 379–383.
62. Jennifer A. Moon. 2013. *Reflection in Learning and Professional Development: Theory and Practice*. Routledge.
63. Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 72–78. Retrieved from http://dl.acm.org/citation.cfm?id=191703
64. Hien Nguyen and Judith Masthoff. 2008. Designing persuasive dialogue systems: Using argumentation with care. In *International Conference on Persuasive Technology*, 201–212. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-68504-3_18
65. Afarin Pirzadeh, Li He, and Erik Stolterman. 2013. Personal Informatics and Reflection: A Critical Examination of the Nature of Reflection. In *CHI '13 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '13), 1979–1988. https://doi.org/10.1145/2468356.2468715
66. David B. Portnoy, Lori AJ Scott-Sheldon, Blair T. Johnson, and Michael P. Carey. 2008. Computer-delivered interventions for health promotion and behavioral risk reduction: a meta-analysis of 75 randomized controlled trials, 1988–2007. *Preventive medicine* 47, 1: 3–16.
67. David Premack. 2004. Is language the key to human intelligence? *Science* 303, 5656: 318–320.
68. Verónica Rivera-Pelayo, Valentin Zacharias, Lars Müller, and Simone Braun. 2012. Applying quantified self approaches to support reflective learning. In *Proceedings of the 2nd international conference on learning analytics and knowledge*, 111–114. Retrieved February 2, 2017 from http://dl.acm.org/citation.cfm?id=2330631
69. Donald A. Schon. 1984. *The reflective practitioner: How professionals think in action*. Basic books. Retrieved from https://books.google.nl/books?hl=en&lr=&id=ceJIWay4-jgC&oi=fnd&pg=PR7&dq=+The+reflective+practitioner:+How+professionals+think+in+action&ots=q80SNZJWqk&sig=BvJKqED7NALlofunXM3jomOp-rc
70. Daniel Schulman, Timothy W. Bickmore, and Candace L. Sidner. 2011. An Intelligent Conversational Agent for Promoting Long-Term Health Behavior Change Using Motivational Interviewing. In *AAAI Spring Symposium: AI and Health Communication*.
71. David W. Schumann, Richard E. Petty, and D. Scott Clemons. 1990. Predicting the Effectiveness of Different Strategies of Advertising Variation: A Test of the Repetition-Variation Hypotheses. *Journal of Consumer Research* 17, 2: 192–202.
72. Shalom H. Schwartz. 1992. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in experimental social psychology*. Elsevier, 1–65.
73. Petr Slovák, Christopher Frauenberger, and Geraldine Fitzpatrick. 2017. Reflective Practicum: A Framework of Sensitising Concepts to Design for Transformative Reflection. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (CHI '17), 2696–2707. https://doi.org/10.1145/3025453.3025516
74. L. Sobell and M. Sobell. 2008. Motivational interviewing strategies and techniques: Rationales and examples. *Retrieved on April* 24: 2015.

75. Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4: 427–437. https://doi.org/10.1016/j.ipm.2009.03.002
76. Yaacov Trope and Nira Liberman. 2010. Construal-level theory of psychological distance. *Psychological review* 117, 2: 440.
77. Gokhan Tur and Renato De Mori. 2011. *Spoken language understanding: Systems for extracting semantic information from speech*. John Wiley & Sons.
78. Harald Walach, Nina Buchheld, Valentin Buttenmüller, Norman Kleinknecht, and Stefan Schmidt. 2006. *Measuring Mindfulness–The Freiburg Mindfulness Inventory (FMI)*. https://doi.org/10.1016/j.paid.2005.11.025
79. Ye-Yi Wang, Li Deng, and Alex Acero. 2005. Spoken language understanding. *IEEE Signal Processing Magazine* 22, 5: 16–31.
80. John R. Ward and Suzanne S. McCotter. 2004. Reflection as a visible outcome for preservice teachers. *Teaching and teacher education* 20, 3: 243–257.
81. Jason D. Williams, Eslam Kamal, Mokhtar Ashour, Hani Amr, Jessica Miller, and Geoff Zweig. 2015. Fast and easy language understanding for dialog systems with Microsoft Language Understanding Intelligent Service (LUIS). In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 159–161.
82. Jason D. Williams, Nobal B. Niraula, Pradeep Dasigi, Aparna Lakshmiratan, Carlos Garcia Jurado Suarez, Mouni Reddy, and Geoff Zweig. 2015. Rapidly scaling dialog systems with interactive learning. In *Natural Language Dialog Systems and Intelligent Assistants*. Springer, 1–13.

Received February 2018; revised April 2018; accepted June 2018