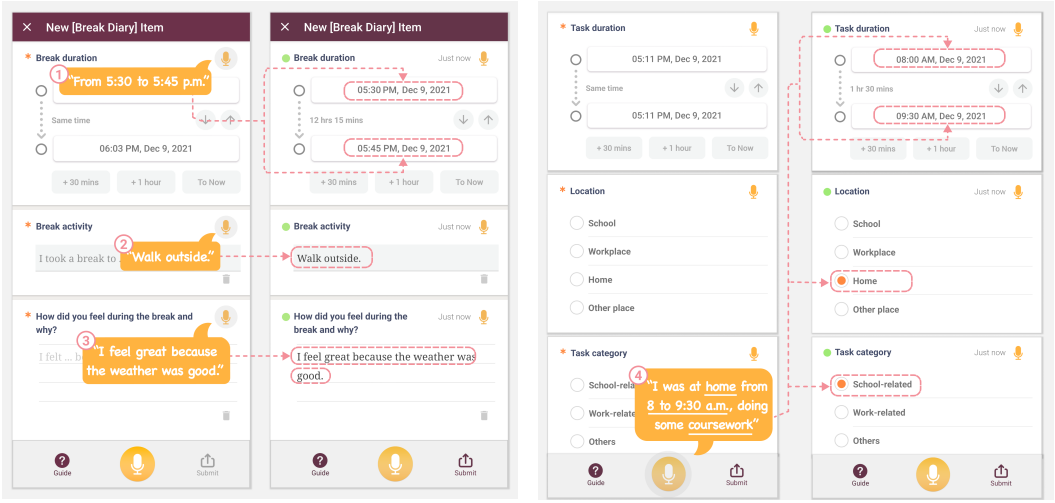
NoteWordy: Investigating Touch and Speech Input on Smartphones for Personal Data Capture

YUHAN LUO*, City University of Hong Kong, Hong Kong SAR
BONGSHIN LEE, Microsoft Research, USA
YOUNG-HO KIM*, NAVER AI Lab, Republic of Korea
EUN KYOUNG CHOE, University of Maryland, College Park, USA



(a) Local speech input (b) Global speech input

Fig. 1. NoteWordy integrates touch and speech input to facilitate the easy capture of personal data on smartphones. With touch input, people can pick time points, select multiple choices, and type text. With speech input, they can capture a single and multiple data fields using ***local speech*** (LS; e.g., ① ② ③) and ***global speech*** (GS; ④), respectively. Please refer to our video figure for interaction details.

Speech as a natural and low-burden input modality has great potential to support personal data capture. However, little is known about how people use speech input, together with traditional touch input, to capture different types of data in self-tracking contexts. In this work, we designed and developed NoteWordy, a multimodal self-tracking application integrating touch and speech input, and deployed it in the context of productivity tracking for two weeks (N = 17). Our participants used the two input modalities differently, depending on the data type as well as personal preferences, error tolerance for speech recognition issues, and social surroundings. Additionally, we found speech input reduced participants' diary entry time and enhanced

*Yuhan Luo and Young-Ho Kim conducted this research while they were at the University of Maryland, College Park.

Authors' addresses: Yuhan Luo, yuhanluo@cityu.edu.hk, City University of Hong Kong, Hong Kong SAR; Bongshin Lee, bongshin@microsoft.com, Microsoft Research, Redmond, WA, USA; Young-Ho Kim, ygho.kim@navercorp.com, NAVER AI Lab, Republic of Korea; Eun Kyoung Choe, choe@umd.edu, University of Maryland, College Park, College Park, MD, USA.

2573-0142/2022/12-ART581 $15.00
https://doi.org/10.1145/3567734

the data richness of the free-form text. Drawing from the findings, we discuss opportunities for supporting efficient personal data capture with multimodal input and implications for improving the user experience with natural language input to capture various self-tracking data.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; **Sound-based input / output**; **Field studies**; **Ubiquitous and mobile computing design and evaluation methods**.

Additional Key Words and Phrases: Self-tracking, personal informatics, speech input, speech interface design, productivity

ACM Reference Format:
Yuhan Luo, Bongshin Lee, Young-Ho Kim, and Eun Kyoung Choe. 2022. NoteWordy: Investigating Touch and Speech Input on Smartphones for Personal Data Capture. *Proc. ACM Hum.-Comput. Interact.* 6, ISS, Article 581 (December 2022), 24 pages. https://doi.org/10.1145/3567734

1 INTRODUCTION

Speech input has been rapidly integrated into our daily life, ranging from speech-to-text services (e.g., live transcription [8], voice typing [38]) to natural language interfaces (NLIs) that respond to user intent (e.g., information query [60], data exploration [33, 66]). Recognizing its fast and flexible nature [12, 59], researchers have started leveraging speech input to facilitate data collection [16, 44, 61, 71]. Particularly in the Human-Computer Interaction (HCI) and Personal Informatics fields, we see a growing incorporation of speech input to capture personal data, such as mood [69], exercise [41], and food practice [37, 40, 64].

However, prior studies incorporating speech input rarely leveraged the capabilities of natural language processing (NLP) to handle the expressiveness of speech. In prior work, speech was used to capture free-form text as an audio file (e.g., Journify [28], FoodScrap [40], Murmur [48]), failing to support the edit of data entry. In other tools, information extraction was possible but in a limited sense—for example, extracting a single type of data (e.g., workout repetition in numbers [41], standard mood measures in Likert scale [13]) or a simple unit of domain-specific items (e.g., food name with quantity [22, 37, 64]). Furthermore, when comparing speech with touch input for personal data capture, speech was often set up separately on a different device (e.g., smart speakers) than touch input on smartphones [41, 64]. But in real-world scenarios, people commonly combine the two modalities on smartphones, because (1) among existing digital devices, smartphones are the most portable and prevalent; and (2) the combination of touch and speech input can complement each other's limitations to facilitate data entry (e.g., compared with touch, speech input is faster for entering text [59], while touch input is more efficient in correcting speech recognition errors [24]).

Within prior work, we have little understanding of how the speech and touch can be used together for different personal data capture. In this light, we are interested in integrating touch and speech input as a whole to capture different types of personal data on smartphones, and aim to answer three research questions:

- **RQ1:** How do people use touch and speech input, individually or together, to capture different types of data for self-tracking purposes?
- **RQ2:** How does the input modality affect people's data capture burden?
- **RQ3:** How does the input modality affect the data richness of free-form text input?

We examined these research questions in the context of productivity tracking because productivity can be characterized by multiple dimensions (e.g., task duration, work output, and mental status) in different data types [29]. We targeted information workers who are also pursuing a graduate degree, who often struggle with maintaining a healthy balance between school and work and thus can benefit from the study [23, 53].

Extending the client app of OmniTrack for Research (O4R) [32], we designed and developed NoteWordy, a multimodal smartphone application integrating touch and speech input for personal data capture (Figure 1). NoteWordy allows people to capture multiple data fields individually or together using their preferred input modality (touch or speech input). In addition to entering one data field at a time with local speech (LS, see Figure 1a), people can enter multiple data fields at once with global speech (GS, see Figure 1b). For example, to capture a Break duration field, people can press on the LS button next to the field and say "*from 5:30 to 5:45 p.m.*" To capture multiple data fields, they can press on the GS button at the bottom and say "*I was at home from 8 to 9:30 a.m., doing some coursework.*" NoteWordy then extracts the information corresponding to "Task duration," "Location," and "Task category" and fills in the corresponding data fields.

Levering O4R's capability in configuring data capture regimens, we created two diaries—"Productivity Diary" and "Break Diary"—consisting of multiple types of data in NoteWordy, and deployed the app to 17 information workers who were also graduate students. From the analysis of participants' data entry patterns with touch and speech input, we found that their modality choices varied by the data type as well as their personal preferences, tolerance for speech recognition errors, and social surroundings. In general, participants liked the convenience of speech input, particularly GS, for capturing multiple data fields at once. They also appreciated touch input, including the fast capture of structured data (e.g., multiple choices, Likert scale) and manual recording of long and complex thoughts. Additionally, we found that speech input reduced participants' time spent on completing the entries in Productivity Diary and enhanced the data richness of free-form text.

This work contributes to: (1) NoteWordy, a mobile user interface that integrates touch and speech input to support capturing multiple types of data; (2) empirical understanding of how people use touch and speech input for self-tracking, including their modality preferences for different data types and how the input modality affected their data capture experience; and (3) implications of designing effective multimodal systems to support personal data capture in various contexts.

2 RELATED WORK

In this section, we describe prior research on personal data collection with speech input and on natural language interfaces (NLIs). We also cover prior studies on collecting productivity-related data and discuss how they informed our study design.

2.1 Personal Data Collection With Speech Input

In recent years, speech interaction has been gaining popularity in a variety of domains [62]. Of particular interests are the self-tracking tools incorporating speech as an input modality [22, 28, 37, 40, 41, 64]. Compared to traditional touch input on smartphones, researchers suggested that speech input is easier for capturing short and structured data, such as workout repetitions [41] and food items [64]. When it comes to capturing long and unstructured data such as food practice, prior work showed that speech input could encourage people to reflect on their thoughts and behaviors through thinking aloud [40, 74]. For example, Luo and colleagues examined how people capture everyday food decisions via free-form audio recording [40]. Through a one-week data collection, the study participants frequently elaborated their responses by providing additional eating contexts, which further facilitated their reflection on current behaviors. In a similar vein, Zhang and colleagues designed Eat4Thoughts that supports people to capture their eating activities via video recording, and found that the audio elements complemented the visual images with richer details about participants' eating experience [74].

However, speech input is not always ideal for capturing personal data. One limitation of speech input is the difficulty with editing when mistakes were made, because people often need to take extra effort to re-record the entire input [40, 41]. Furthermore, there is a general impression that

speech interface is vulnerable to recognition errors—once an error occurs, it can be difficult to recover [41, 64]. As a result, people who have negative experience using speech input may resist using it again. Another limitation is related to capturing personal data in public spaces: some people feel embarrassed talking to their phone in front of others [40, 41] or concerned about their privacy being disclosed [40, 51].

Unlike the abundance of research on speech's potentials and limitations as an input modality for personal data capture, little research has examined how speech can work with touch input on the same device to capture different types of data. In this work, we seek to support *multimodal self-tracking* by integrating speech and touch input on smartphones, and examine how people use the two modalities in a context where multiple types of data need to be collected. Moreover, we reveal how touch and speech affected data capture burden (e.g., time spent) and data richness (e.g., level of details) with both quantitative and qualitative understandings.

2.2 Natural Language Interfaces (NLIs) for Data Capture

Advancements in natural language processing (NLP) have accelerated the rise of Natural Language Interfaces (NLIs), in which words, phrases, or sentences perform as commands for creating, selecting, and modifying data [20, 26, 50]. To interpret and execute these commands, NLIs employ both rule-based approaches and machine learning techniques, which aid in processing the unstructured language sources into structured objects such as entity, time, and event [21, 54, 56]. A typical example is reminder settings with voice assistants (e.g., Amazon Alexa [7]): people can say "*remind me at 8 am every day to exercise*," from which the service automatically extracts the time, event, and whether this is a recurring reminder. Due to the complicated sentiment and ambiguity in natural languages, current NLIs are not yet generalizable to handle all kinds of input [19]. Even the state-of-the-art NLP systems rely on a knowledge base (e.g., vocabularies, syntax rules) or domain-specific training data [21, 56].

Among self-tracking applications that incorporate speech input, only a few of them are equipped with NLIs to process natural language input [22, 37, 41, 64]. For example, Luo and Colleagues developed TandemTrack, an exercise tracking system integrating a mobile app and a smart speaker application, which allows people to capture their exercise repetitions by speaking to the smart speaker (e.g., "*I did 20 push-ups*") [41]. In the domain of food tracking, Silva and colleagues implemented a multimodal food journaling app called ModEat, which supports food recognition from text, database search, barcode, and speech input [64]. Leveraging an external NLP service, ModEat can recognize a variety of foods and calculate people's calorie consumption. Similarly, Korpusik and colleagues built their own NLP solution and developed Coco Nutrition, a conversational calorie counter that detects food items and quantities from both text and speech input [37]. In addition, the commercial app Talk-to-Track can extract food items or exercise activities from spoken language, but requires people to deliberately separate each item by saying "comma" [22], which limits the flexibility of speech input.

While demonstrating the promises of NLIs to support personal data capture, existing work predominantly focused on capturing only a single type of data such as numbers or a single unit of domain-specific items (with food items—name and quantity—being the most common). However, people often track multiple data about their target activities in different types (e.g., time, location, activity type, and other contexts) [31], which is not well-supported by existing tools. In this work, we set out to realize the benefits of flexible speech input for self-tracking with NoteWordy. By extracting multiple data values from a single utterance, NoteWordy allows people not only to take advantage of speech input's fast and flexible data capture, but also to easily review the extracted information from their data.

2.3 Productivity of Working Graduate Students

We identified productivity tracking as our study context because it allows us to create a data capture regimen that consists of multiple structured and unstructured data, which is essential to answer our research questions. As a multifaceted concept, productivity can be characterized by several aspects of people's work and life, such as work product [29], time spent on tasks and breaks [18, 27], self-rated productivity score [45, 47], as well as physical and mental status [29, 42, 58]. These data can be captured in different types, including timespan, multiple choice, Likert scale, and free-form text. Due to the subjective nature of the data (e.g., self-rated productivity score [45, 47]), prior work predominately employed touch-only input that requires people to perform a series of manual selections and typing [11, 18, 29, 45, 47].

We targeted working graduate students—information workers who are pursuing a graduate degree full-time or part time, because they often juggle multiple tasks and face challenges in maintaining a healthy balance between school and work [9, 15, 23, 49, 53]. Unlike undergraduate students who usually have structured course schedules with GPA-oriented goals [63], graduate students tend to have more flexible schedule but may experience more stress due to career transition, financial burdens, or family obligations [55]. Therefore, time management can be more challenging and complicated for graduate students, especially those who are also employed for another job. To understand how working graduate students move between different roles, prior research found that segmenting time spent on school and work tasks significantly improved one's ability to handle conflicts and mental stress [17]. Thus, we posit that working graduate students could benefit from tracking their productivity data. Furthermore, the data collected in the study can serve as ecologically valid sources that help researchers understand working graduate students' productivity-related behaviors and inform the design of productivity tools for this particular group [15, 49, 53].

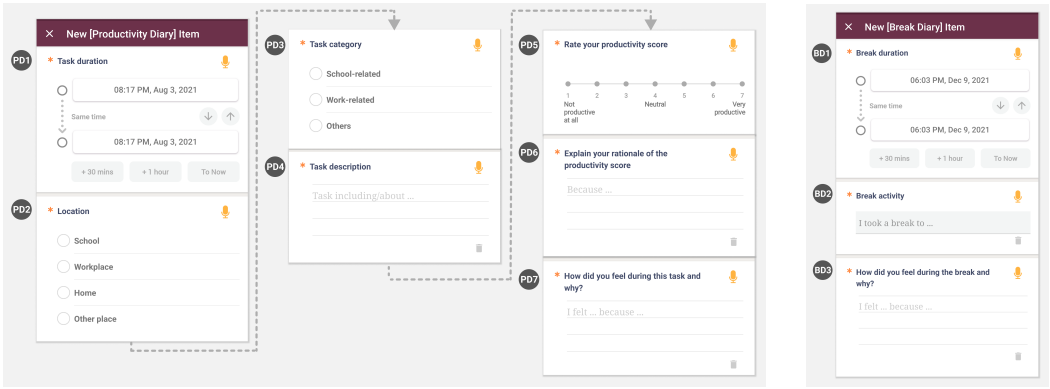
3 NOTEWORDY

NoteWordy was built upon the client app of OmniTrack for Research (O4R) [32], which already supports capturing different types of data with touch input. Our design and implementation of NoteWordy thus focused on incorporating speech input to capture individual and multiple data fields. In the following, we first describe our design rationales, and then present NoteWordy's speech interface along with implementation details.

3.1 Design Rationale

3.1.1 DR1: Provide Both Touch & Speech Input Capabilities. People have individualized preferences for the input modality [41] while their choices also being affected by external factors such as social environments [40, 41]. To examine how people choose between or combine touch and speech input to capture different types of data (RQ1), we provided both touch and speech input options for all the data fields, instead of designating speech input for certain data fields only.

3.1.2 DR2: Enable Flexible Data Capture Through Natural Language Input. As a natural input, speech offers two advantages. First, it allows people to capture the same data with different expressions [33, 35]. For example, people can capture time points in standard or (e.g., "*8 in the morning*") relative (e.g., "*two hours ago*") forms. Second, people can capture multiple data fields in one utterance without following a particular order. In the example illustrated in Figure 1a, "Break duration," "Break activity," and "How did you feel during the break and why" can also be captured in one sentence, such as "*I walked outside from 5:30 to 5:45 p.m., feeling great because the weather was good*," instead of individually saying "*from 5:30 to 5:45 p.m.*," "*walk outside*," and "*I feel great ...*" To maximize the advantages of speech input, we enabled the capturing of multiple data in a single utterance, in addition to the individual capturing of each data field.



(a) Productivity Diary (b) Break Diary

Fig. 2. The questions asked in Productivity Diary (a): task duration (PD1), location (PD2), task category (PD3) and description (PD4), productivity score (PD5) and rationale (PD6), and feelings during the task (PD7); and in Break Diary (b): break duration (BD1), break activity (BD2), and feelings during the break (BD3).

3.1.3 DR3: Design for Clear Speech Guidance. One main challenge that people often face with speech interface is the discoverability (i.e., the ability to discover the correct speech commands to interact with the system) [34]. Without clear guidance, people are unsure about how to phrase an utterance [41, 67], and may end up abandoning speech input and turning to other modalities (e.g., touch) [41]. To help people understand the capabilities and limitations of NoteWordy, we guided them through the input process by providing utterance examples when the speech input is initiated. Furthermore, the current speech recognition and data processing techniques are not perfect due to the complexity of natural language [35] and various external noises (e.g., background sounds, microphone quality) [72]. Thus, when a speech recognition error occurs, it is important to acknowledge the error and remind people of the input hint with examples.

3.2 Data Capture With NoteWordy

3.2.1 Diary Design: Productivity Diary & Break Diary. Drawing from prior research on productivity data collection, we focused on three aspects that play important parts in one's daily productivity: tasks [29], breaks [18], and mental status (e.g., feelings) [58]. We configured two diaries in NoteWordy[1]: Productivity Diary and Break Diary. In Productivity Diary (Figure 2a), people are asked to answer questions about each task. In addition to entering task duration (PD1), task location (PD2), task category (PD3), and detailed task description (PD4), participants need to rate their productivity score in a standard Likert scale (PD5) [45, 47], explain the rationale of the rating (PD6), and describe how they felt during the task and why (PD7). In Break Diary (Figure 2b), we shortened the questions to focus on people's break duration (BD1), break activity (BD2), and how they felt during the break and why (BD3).

3.2.2 Local Speech (LS) Input. To provide both touch and speech input capabilities (DR1), we placed a ***local speech*** (LS) button on each data field (Figure 1a). With the "push-to-talk" operation, the system records the speech input while people are pressing on one of the LS buttons and extracts the information related to the field that the button is being pressed. The system handles natural

[1]NoteWordy allows researchers to design their data capture regimens on a web-based dashboard. Based on the configuration of the data capture regimens, the server populates corresponding diaries to a new participant's account when they first sign in [32].

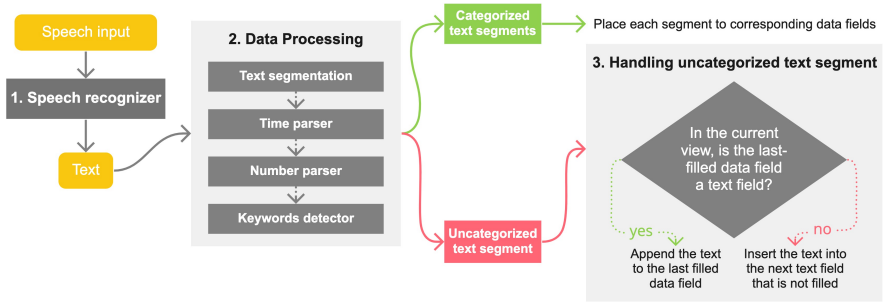


Fig. 3. The pipeline that processes GS input: 1. Transcribing speech input into text; 2. Extracting the content data related to the structured data fields from the text and categorizing other text segments based on keywords; 3. Handling uncategorized text segment.

language input (DR2), allowing people to record the data in the ways they like. In PD1, for example, people can provide the start and end time or mention the duration with only one of the two points. To capture productivity score in PD5, people can say a number from one to seven or a label from "*not productive at all*" to "*very productive*" (e.g., say "*productive*" referring to "*6*"). When people press on the LS button on any of the text fields (PD4, PD6, PD7), all the transcribed text from their speech input will be entered into that field. They can also append more text to that field by pressing on the LS button and speaking again.

3.2.3 Global Speech (GS) Input. We provide a ***global speech*** (GS) button that is unattached to any data fields (Figure 1b) so that people can capture multiple data fields at once (DR2). GS also adopts the "push-to-talk" operation that records speech input while people are pressing on the GS button. People are asked to include certain keywords in their utterances to help the system extract the key information. The recommended keywords for each text field are displayed in gray text as

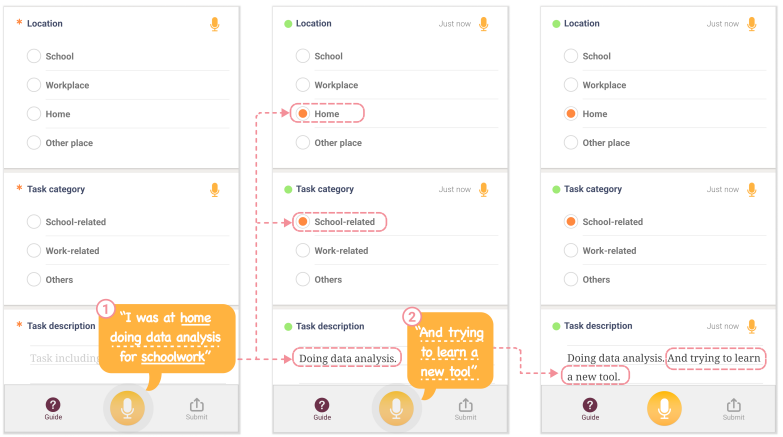
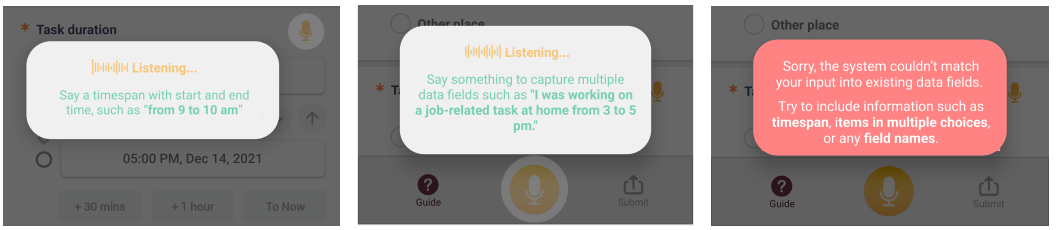


Fig. 4. An example of how GS handles uncategorized text in Productivity Diary: if no data field in the current view is filled or the last filled data field is not a text field, the uncategorized text will be inserted into to the next text field coming along ①; if the last filled data field is a text field, the uncategorized text will be appended to that text field, allowing people to incrementally add information to the same text field with GS ②. (The keywords that helped the system extract the information are underscored with solid lines and the unrecognized text is underscored with dotted lines).



(a) (b) (c)

Fig. 5. The speech input dialog that displays contextual messages to guide people through the recognition process: (a) when the LS button on "Task duration" is pressed; (b) when the GS button is pressed with "Task duration," "Location," and "Task category" all visible; (c) when the system fails to recognize the GS input.

a hint (e.g., "*Tasks including/about ...*" under PD4 (Figure 2)). Figure 3 illustrates the pipeline that processes GS input.

When a text segment neither belongs to a structured data field nor contains the keywords of existing text fields, NoteWordy handles it in two ways: (1) in the current view, if no data field is filled or the last filled data field is not a text field, the segment will be inserted into the next text field coming along (Figure 4 ①); (2) if the last filled data field in the current view is a text field, the uncategorized segment will be appended to that text field, allowing people to incrementally add information to the same text field with GS (Figure 4 ②). This design assumed that people are likely to complete the diary entries by following the order of the questions and may need to enter data into the same text field multiple times. if uncategorized segment is placed to the wrong text field, people can delete the text by pressing on the clear button at the right bottom of that field.

3.2.4 Contextual Guide & Error Feedback. When people press on the GS button or any of the LS buttons, a speech input dialog pops up to guide them through the recognition process while dimming the screen behind (see Figure 5). Before people start talking, the dialog displays a contextual message explaining what they can say with an utterance example, which is based on the data field that the button is placed on (e.g., showing "*from 9 to 10 am*" for the "Task duration" field (Figure 5a)). When the GS button is pressed, the dialog displays an utterance example based on the data fields that are visible in the current view (e.g., showing "*I was working on a job-related task at home from 3 to 5 p.m.*" when Task duration, Location, and Task category fields are all on the screen (Figure 5b)). While people are talking to NoteWordy, the dialog displays live transcripts of the speech input so that they are aware of how their utterance is being recognized. This also prevents people from releasing their finger before the system completes recognition. If NoteWordy fails to recognize the speech input, an error message pops up to inform people of what might be wrong and suggests alternative utterances that they can try (Figure 5c).

3.3 Implementation

Extending O4R, NoteWordy is written in Kotlin [2] on Android platform. We used Microsoft Cognitive Services [4] as a speech-to-text recognizer because (1) it allows developers to customize timeout for continuous recording, so that we could avoid problems caused by pauses in the middle of recording; and (2) the service provides automatic punctuation, which could help with text segmentation. We used SmileNLP [6], a machine learning engine to further segment the utterances and to handle different forms of the same word (e.g., "*feeling*" and "*felt*" are different forms of "*feel*"). We also incorporated Natty [5], a time parser, to process different time expressions. To improve the recognition accuracy, we appended a set of keywords related to the study context (e.g., "*coursework*" and "*schoolwork*" are synonyms for "*school-related*") to NoteWordy's vocabulary.

NoteWordy collects the transcribed text of all the user utterances, but not the original audio recordings. The data are securely stored on a virtual machine hosted on the university's server. People can access their data in the app by revisiting the raw entries or aggregated visualizations (e.g., the number of daily entries, productivity score across time). The details of the visualization design are described in [31].

4 METHODS

We deployed NoteWordy remotely to 17 working graduate students for two weeks, followed by debriefing interviews. The study was approved by the university's Institutional Review Board (IRB).

4.1 Participants

We advertised the study through the university mailing-list, Reddit (under the subreddit "r/GraduateSchool" and "r/MBA"), and Facebook (under the group of "Graduate School"). We also approached several Reddit users who posted discussions about time management as working graduate students,

Table 1. Participants' demographic, student status, majors, employment types, work modes, and experience with speech interface.

ID	Age	Gender	Location	Student type	Major	Off-campus occupation	Employee type	Work mode	Experience with speech interface
P1	27	M	CA	Full-time master	MBA	UX Designer	Full-time	Remote	Neutral
P2	30	M	MI	Part-time master	Data Science	IT Administrator	Full-time	Remote	Positive
P3	30	F	MI	Part-time master	Data Science	Data Analyst	Full-time	Remote	Positive
P4	24	M	MD	Part-time master	HCI	Data Engineer	Full-time	Remote	Positive
P5	25	F	MD	Full-time Ph.D.	Computer Science	Art Designer	Freelancer	Remote	Neutral
P6	35	M	MD	Part-time master	Statistics	Research Analyst	Full-time	Hybrid	Positive
P7	26	M	NY	Part-time master	MBA	Photographer	Part-time	Hybrid	Positive
P8	30	F	MD	Full-time master	Library Science	Music tutor	Part-time	Hybrid	Neutral
P9	24	F	MI	Full-time master	Medicine	Behavior technician	Part-time	In-person	Positive
P10	22	M	MD	Full-time master	Computer Science	Researcher	Part-time	In-person	Positive
P11	35	M	WA	Full-time master	Theoretical Physics	Database Operator	Part-time	Hybrid	Positive
P12	37	F	MD	Part-time master	Library Science	Research Analyst	Full-time	Remote	Neutral
P13	25	M	MD	Full-time master	HCI	Newsletter Coordinator	Part-time	Remote	Neutral
P14	26	F	MD	Full-time master	Classics	ESL Instructor	Part-time	Hybrid	Positive
P15	28	F	VA	Part-time Ph.D.	Aerospace Engineering	Aerospace Engineer	Full-time	Remote	Positive
P16	29	M	MD	Full-time Ph.D.	Computer Science	Researcher	Part-time	In-person	Neutral
P17	37	F	OH	Full-time Ph.D.	Social Science	Researcher	Full-time	Hybrid	Neutral

and asked if they would like to participate in the study. We selected participants using a screening questionnaire, and our inclusion criteria were adults who (1) are fluent in English; (2) possess an Android mobile phone with an OS version 4.4 or above; (3) are enrolled in a graduate program at a university (master's or Ph.D. level); (4) are employed full-time or part-time outside the university, working in front of computers at least 20 hours per week in addition to schoolwork; (5) are interested in time management and curious about their time spent between school and work; (6) have no visual, motor, or speech impairments; (7) have experience using speech interfaces and are willing to use it daily; (8) have stable access to the Internet; and (9) have a computer with a webcam, microphone, and speaker so that they can communicate with the researchers via video chat. By looking for graduate students who worked off-campus, we excluded graduate research or teaching assistants, because their work is often a part of or overlaps with their schoolwork and there may not be a clear boundary between their work-related and school-related tasks.

Among the 66 qualified individuals we contacted, 27 replied to us, 20 completed the study tutorial, and 17 completed the data collection. Others dropped out due to technical issues or became unresponsive as the study progressed. As Table 1 shows, our 17 participants' (8 female) age ranged from 22 to 37 (*Median* = 28, *SD* = 4.7) and lived in different regions in the US. They majored in different fields of study with different jobs. All the participants had used speech interface (e.g., voice assistant on their phones, smart speaker) before with positive or neutral experiences. Three participants (P2, P3, P17) reported themselves as working student-parents, and two (P2, P6) reported that they had been diagnosed with attention-deficit hyperactivity disorder (ADHD).

4.2 Study Procedure

The study consisted of three stages: (1) tutorial, (2) two-week data collection, and (3) debriefing interviews (optional). After completing the data collection, each participant received a $30 Amazon gift card as compensation. Those who opted in to the debriefing interviews received an additional $10 gift card.

4.2.1 Tutorial. At the beginning of the study, we had a one-on-one remote tutorial with each participant (45 minutes). We asked participants to share their phone screen with us via TeamViewer QuickSupport [68] so that we could instruct them to install NoteWordy and watch them interact with the app in real-time. We then shared our computer screen via Zoom so that participants can see how their phone screen was being displayed to us. Prior to screen sharing, we asked participants to remove any sensitive information from their home screen and turn off incoming notifications to mitigate the risks of accidental privacy disclosures.

During the tutorial, we went through the study procedure and described the types of data that participants needed to collect. After demonstrating how to enter each data field with LS and GS, we led a short practice session with each participant. First, we asked the participant to enter the data fields, individually with LS and together with GS, by following the example utterances we prepared. Next, the participant freely explored the touch and speech input to get familiar with the interface for about three minutes. The participant then needed to think about a recent task and a break, and complete one entry respectively in Productivity Diary and Break Diary. Lastly, we explained that NoteWordy's speech recognition was not perfect (e.g., missing keywords) and situations where recognition issues might happen (e.g., talking too far away from the microphone).

4.2.2 Data Collection. The data collection started the next day after the tutorial and lasted for two weeks, during which participants used NoteWordy to capture their tasks and breaks. To ensure that participants capture their tasks across different times during the day, we segmented the daytime into four windows: 9 to 12 p.m., 12 to 3 p.m., 3 to 6 p.m., and 6 to 9 p.m. As a minimal requirement of data capture in Productivity Diary, each participant needed to capture one task in three of the

above time windows per day (e.g., one task from 9 to 11 a.m., one task from 4 to 6 p.m., and another task from 6 to 7 p.m.). In Break Diary, participants needed to capture at least one entry per day. Our study focused on capturing "intentional breaks" that participants took to refresh and relax instead of "unintentional breaks," such as being distracted by social media or going to the bathroom. To help participants remember to capture their data on time, we configured four reminders in NoteWordy, which were sent at 12 pm, 3 pm, 6 pm, and 9 pm respectively on daily basis. During the 14 days, each participant was allowed to skip their daily entries for two days. When a participant had already skipped two days' entries, we would send them an email notifying that they cannot skip any days before the end of the study. Those who could not meet the minimal data capture requirement were dropped from the study.

4.2.3 Debriefing Interviews. Upon the completion of data collection, we contacted each participant to ask if they were interested in attending a debriefing interview for 30 to 45 minutes via Zoom. To help participants better recall their study experience, we asked them to open their diary entries on NoteWordy and share their phone screen with us using TeamViewer QuickSupport. All the participants opted in to do the interview, during which we asked questions about how they chose between or combined touch and speech input in different scenarios, their preferences for LS, GS, and touch input, and the challenges they faced in completing their diary entries.

4.3 Data Analysis

Our study generated a mix of quantitative and qualitative data, including participants' interaction logs with NoteWordy, their diary entries, and the interviews on their study experience. Here, we describe how we analyzed these data to answer our research questions.

4.3.1 Log Data Analysis. We first summarized the descriptive results of participants' diary entries and the input modalities that they used to capture each data field. We then looked into whether the use of speech input reduces participants' time spent on completing the diary entries (i.e., the duration between when the participant opened an entry and when they submitted the entry). To take individual differences into account, we used multilevel linear regression modeling by treating the use of speech input as a fixed effect and participant as a random effect [2]. To further investigate how participants used touch, GS, and LS input, we summarized their data input patterns by grouping the data fields that were typically captured together. We also broke down the usage of the input modalities by each participant to examine the individual variations in modality preferences.

4.3.2 Diary Entry Analysis. To examine whether and how input modality influence the data richness of unstructured (i.e., free-form text) input, we analyzed the responses to the three text fields in Productivity Diary: PD4—Task description, PD6—Productivity rationale, PD7—Feelings and the two text fields in Break Diary: BD2—Break activity, BD3—Feelings . First, three researchers independently analyzed a subset of the 3860 text field responses (811, 21%) and created a set of labels characterizing the richness of the responses. Following prior work on analyzing the data richness of open-ended responses [40, 57, 65], we refined the coding scheme through rounds of comparison and discussions, and agreed to categorize the responses according to whether they *answered the question with specificity* and whether they were elaborated with *additional contexts to help researchers better understand the situation*. We categorized three types of responses: *generality*, *specifics*, and *specifics with additional contexts* [3] (See Table 5 for details). Based on the coding scheme,

[2]Based on a-priori power analysis, each regression model in this paper reached over 90% power (α = .05, β = .20) with a medium effect size (*Cohen's* f^2 > .15).

[3]We initially coded different types of contexts that the responses mentioned (e.g., other people, task procedure, prior work experience), but did not find prominent themes from these contexts.

two researchers revisited the same subset of data and separately coded them, reaching near-perfect agreement (Cohen's κ = .84). After resolving the discrepancies, the first author coded the remaining responses. Next, we used multinomial logistic regression to examine if input modality tended to affect the data richness of each text field, while treating participant as a random effect.

4.3.3 Interview Data Analysis. We audio recorded all the interviews and transcribed them into text. Three researchers separately analyzed the transcripts and built an initial list of codes starting with a top-down (deductive) approach to identify factors influencing participants' modality choice (e.g., *data types*, *environmental constraints*). After several iterations of coding, we organized our codes into emerging themes using bottom-up (inductive) thematic analysis [10] to characterize the advantages and limitations of the two modalities.

5 RESULTS

Throughout the two weeks, NoteWordy collected 1032 entries in Productivity Diary (60.7 entries per participant) and 382 entries in Break Diary (22.4 entries per participant). As Table 2 shows, 43.4% of the diary entries were completed by touch-only input, 12% were completed by speech-only input [4], and the remaining 44.7% were completed with some data fields filled by touch input and others filled by speech input (touch + speech). In this section, we present participants' usage of touch and speech input (RQ1), and how the input modality affected their data capture burden (RQ2) and data richness of free-form text fields (RQ3).

5.1 RQ1: Usage of Input Modalities

Table 3 summarizes participants' input patterns contributing to the data fields that were typically captured together, including a combination of structured (i.e., timespan, multiple choice, Likert scale) and unstructured data (i.e., text). Here, we describe four prominent input patterns.

5.1.1 Modality Choice By Data Type. We found that touch input was most frequently used for capturing structured data including timespan, multiple choice, and Likert scale questions (**T1, TS1**), because the interaction was "*easy*" and "*familiar*" to our participants. In particular, timespan as a structured data was also frequently captured by speech input (GS or LS) in 32.7% of the Productivity Diary entries (**TS1, S1**) and 43.2% of the Break Diary entries (**S2**). Participants found speech input more effective than touch input in capturing timespan, because "*manually selecting when it started and ended is tedious, so I just went to the little mic and clicked on it, and say '9 to 12 p.m. yesterday' and found it very easy*" (P7). Speech input also allow participants to describe their task and break duration in different ways, including providing the standard start and end times (e.g., "*8 to 9:30 p.m.*"), relative time points (e.g., "*started 3 hours ago till now*"), or specific duration (e.g., "*started at noon and lasted 45 minutes*"). In addition, participants highlighted the convenience of capturing

Table 2. The number of entries that were completed by touch, speech, and speech plus touch input.

Input modalities	Total	Productivity Diary	Break Diary
Touch input only	613 (43.3%)	429 (41.5%)	184 (48.2%)
Speech input only	169 (12.0%)	38 (3.7%)	131 (34.3%)
Speech + Touch input	632 (44.7%)	565 (54.7%)	67 (17.5%)
Total	1414	1032	382

[4]We use "speech-only input" to denote people using LS or GS input to enter their data, although it requires touching the speech button (i.e., the "push-to-talk" operation).

Table 3. Summary of input patterns contributing to data fields that were typically captured together. The modality column indicates the input modalities that were used (T: touch, GS: Global Speech, LS: Local speech; the percentage of input patterns was calculated based on the combinations of field types, not the total entries).

Data fields	Input pattern	Modality	Freq	Example
Structured data				
Timespan & multiple choices [a] **(*n* = 1032)**	**T1.** T <timespan> + T <multiple choices>	T	660 (63.9%)	Pick start & end time in task duration then Select the location and task category
	TS1. GS/LS <timespan> + T <multiple choices>	T GS LS	141 (13.7%)	or task duration say " *8 to 10 a.m.* " then Select the location and task category
	S1. GS <timespan & multiple choice>	T	196 (19.0%)	say " *work task* *at home* *from* *9 to 12 p.m.* "
	Miscellaneous [f]		35 (3.4%)	
Structured data + Unstructured data				
Timespan & text fields [b] **(*n* = 382)**	**T2.** T <timespan> + T<text>	T	184 (48.2%)	Pick start & end time in break duration then type break activity and break duration
	S2. GS <timespan & text>	GS	165 (43.2%)	say "*I* *walked outside* *from* *4 to 4:30 p.m.* , *feeling* *refreshed because the weather was nice* "
	Miscellaneous		33 (8.6%)	
Multiple choices & text field [c] **(*n* = 1032)**	**T3.** T <multiple choices> + T<text>	T	473 (45.8%)	Select task category then type task description
	TS2. T <multiple choices> + LS <text>	T LS	343 (33.2%)	Select task category then task description say "*writing a report for my class*"
	S3. GS <multiple choice & text>	GS	165 (16.0%)	say " *School-related* *tasks* *on python codes* "
	Miscellaneous		51 (5.0%)	
Likert scale & text field [d] **(*n* = 1032)**	**T4.** T <Likert scale> + T<text>	T	434 (42.1%)	Pick productivity score then type productivity rationale
	TS3. T <Likert scale> + GS/LS <text>	T GS LS	428 (41.5%)	Pick productivity score then productivity rationale say "*Got most work done fast*"
	S4. GS <Likert scale & text>	GS	149 (14.4%)	say "*I was* *somewhat productive* *because I completed the task but it ended up taking more time than planned* "
	Miscellaneous		21 (2.0%)	
Unstructured data				
Multiple text fields [e] **(*n* = 1414)**	**T5.** T <each text field>	T	619 (43.8%)	Type productivity rationale and feelings
	S5. LS <each text field>	LS	390 (27.6%)	productivity rationale say "*I wasn't very focused*" then feelings say "*tired since I did a lot of chores*"
	S6. GS <all text fields>	GS	379 (26.8%)	say "*I* *had some snacks* *and* *felt satisfied because those are my favorites* "
	Miscellaneous		26 (1.8%)	

[a] Productivity Diary: task duration (PD1) & location (PD2) & task category (PD3).
[b] Break Diary: break duration (BD1) & break activity (BD2) & feelings (BD3).
[c] Productivity Diary: task category (PD3) & task description (PD4).
[d] Productivity Diary: productivity score (PD5) & productivity rationale (PD6).
[e] Productivity Diary: productivity rationale (PD6) & feelings (PD7); Break Diary: break activity (BD2) & feelings (BD3).
[f] Miscellaneous: instances that are not categorized under the prominent input patterns (e.g., T/LS <timespan> + GS/T <multiple choices>).

free-form text using LS and GS: "*I probably would never manually input the open ended questions unless I really had to, because it would just take too much time to type the details*"(P14).

5.1.2 Starting With GS, Followed By LS or Touch. Six participants in our study (P7, P9, P11, P13, P14, P17) showed a strong preference for GS because it was "*faster*," "*intuitive*," and "*more accurate than expected*." Oftentimes, they started capturing multiple data together and then adjusted individual data fields as needed (**S1**–**4**, **S6**): "*I started off with the global speech. For the most part, it did a good job capturing what I was saying. Sometimes there will be just spelling errors, so I would make manual adjustments*"(P9). It was noteworthy that although participants rarely used LS to individually capture multiple choice and Likert scale questions, they used GS to capture these two types of data together with other fields (**S1, S3–4**), because GS saved their effort to "*click and hold for every single field*" (P10). When asked to compare between LS and GS, P7 remarked that LS was like "*individual voice commands*" and GS was more "*close to natural language*."

5.1.3 GS Usage in Productivity Diary vs. Break Diary. Interestingly, we found that in Productivity Diary, GS was used in less than 20% of entries (**S1**, **S3–5**); while in Break diary, GS was used in 43.2% of the entries (**S2**). Participants found GS most useful when they could "*naturally link multiple data in one sentence*", but in Productivity Diary, it was not always as natural to do so: "*Sometimes I will try to say things like 'I worked on school-related things at school' or 'working on a work-related task at workplace,' which for me sounds a little awkward to say*" (P11). Nine out of 17 participants explained that they preferred using GS in Break Diary, because the diary was shorter and all the data fields were visible on the screen at once, allowing them to quickly skim what information to capture and speak without scrolling: "*I could see everything on the screen at same time, so I didn't have to worry that I was going to miss a question or something like that*" (P14).

5.1.4 Variations in Modality Preferences. Figure 6 illustrates the modality usage of each participant, showing large variations in modality preferences across individuals: seven participants (P1, P7, P9, P14, P13, P2, P6) used speech input in more than 50% of their diary entries; four participants (P10, P11, P17, P15) used speech input occasionally, but less often than touch input; and the remaining six participants (P4, P12, P3, P5, P16, P8) used touch input most of time, with fewer than 25% of diary entries involving speech input.

Participants who preferred touch input tended to enter multiple choices and Likert scale by quickly tapping the screen; they also valued manual typing as an important way to capture free-form text in many scenarios. Due to ***privacy concerns***, six participants (P6, P11, P12, P15, P16, P17) did not want their work-related information to be overheard by others in offices. Five participants (P3, P4, P5, P10, P12) pointed out that capturing personal data by talking to their phones was an ***"atypical (social) norm"***. Rather than worrying about privacy, they worried about"*over-sharing*" their life that others did not care about, and felt more comfortable using touch input as a habit. Five participants (P3, P5, P8, P16, P17) found themselves "***better at writing than speaking***" when describing complicated thoughts: "*I think the rationale of productivity score and feelings about the tasks had a little more involvement. I guess for me, it's just easier to write than speaking out.*"

Participants who preferred speech input expressed their excitement about the ***convenience*** and ***accuracy*** of LS and GS: "*I would say it's pretty accurate. The global speech is very impressive, you don't really need to remember the keywords specifically, because as long as you follow the diary, it catches what you are trying to say*" (P7). They also enjoyed "*thinking out loud*" with speech input, because it was easier for **in-situ data capture**, as P2 explained: "*I actually liked speech a lot, especially when I recorded a task that just happened, because I might need to wrap it up so I don't really want to type, but I also remembered everything so it's kind easy to say it aloud.*"

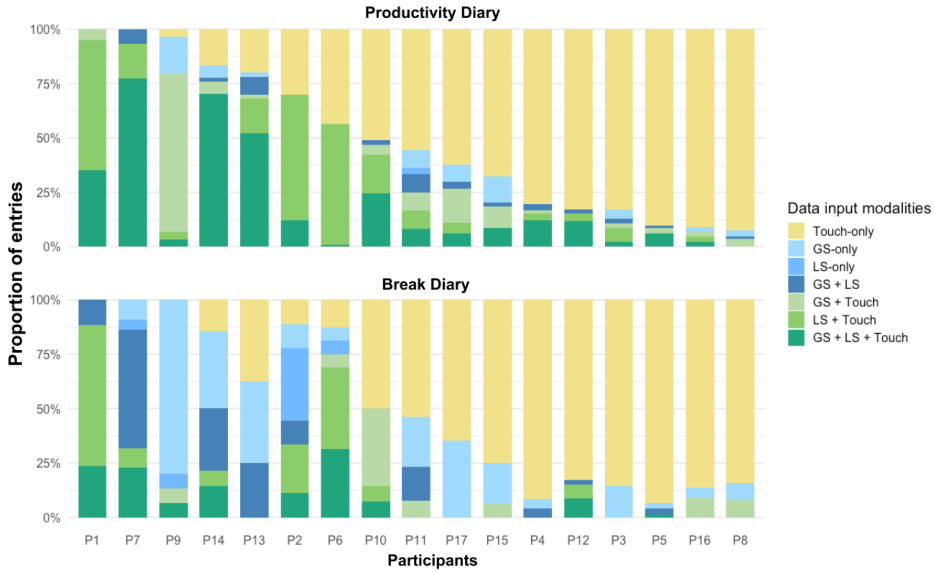


Fig. 6. The usage of speech (LS and GS) and touch input in completing the diary entries per participant (participant id is ordered by the proportion of speech input usage in Productivity Diary).

5.2 RQ2: Data Capture Burden

We report participants' data capture burden with NoteWordy from four aspects: (1) time spent on diary entry; (2) adoption barriers of GS; (3) burden associated with speech recognition; and (4) data mismatch issues.

5.2.1 Time Spent on Diary Entry. On average, participants spent 143.7 seconds per entry in Productivity Diary and 78.4 seconds per entry in Break Diary (See Table 4). In both diaries, speech-only and speech + touch input entries took less time to complete than touch-only entries, and the use of speech input significantly reduced the time spent on completing the entries in Productivity Diary (b = -.38, p = .004).

5.2.2 Adoption Barriers of GS. While acknowledging the intuitiveness of GS, P11, P12, P13, and P15 considered it a new interaction paradigm, and therefore avoided using GS at the beginning: "*I wasn't really sure about the commands for the global ... So just for my own reliability sake, I was typing it or use the individual ones (LS)*" (P13). While P13 eventually became comfortable using GS later after some 'trial and error,' others still stayed with touch and LS input due to unfamiliarity. In addition, six participants (P1, P2, P3, P4, P6, P12) reported that GS could cost extra mental load. For example, P3 noted that "*although the global speech was really really cool, I found myself not*

Table 4. Participants' time spent (seconds) on completing the diary entries using speech, touch, and both input modalities.

Entry type	Productivity Diary	Break Diary
Touch-only entries	175.9	86.7
Speech-only entries	115.9	65.5
Touch + Speech entries	121.1	81.0
Average	143.7	78.4

ready to use it ... I didn't always have all my thoughts together of exactly what I wanted to say for every single part. I would forget what else needed to be said so I'd have to stop and think." Sometimes, participants provided long responses to describe their tasks, productivity rationale, and feelings. In these cases, they preferred LS rather than GS for capturing details in each field: "*I always provide as many details as possible for the study. If I use the global, I would add more details with the local mic (LS) anyway, so I did not use it very often*" (P6).

5.2.3 Speech Recognition Issues. The invalid utterances (from both LS and GS input) logged by NoteWordy revealed several data capture burdens associated with speech recognition issues. Of the 93 invalid utterances, 77 (82.8%) were related to ***number recognition***, which caused errors in timespan (e.g., "*started five*" being recognized as "*started fine.*") and Likert scale fields (e.g., "*productivity: two*" being recognized as "*productivity: to.*"). Learning from that experience, some participants were able to avoid recognition issues by adding more phrases in their utterances or eliminating specific words. For example, P7, P14, and P17 found that mentioning time-related phrases such as "*a.m.*" and "*in the morning*" could improve the accuracy of time recognition. After realizing that the word "*to*" was often recognized to the number "*two*," P15 decided to use the word "*till*" to describe timespan (e.g., "*7 till 9*"). To enter productivity score, participants used the label (102, 68.5%) more frequently than the number (47, 31.5%) (e.g., saying "*very productive*" instead of "*7*") as a strategy to prevent the system from missing their productivity score (P11, P17).

While not explicitly logged by NoteWordy, we noticed other types of recognition issues. For example, in entering productivity score (PD5), utterances such as "*relatively productive*" were interpreted as "*productive*" (*6*), even though they were more likely to indicate "*somewhat productive*" (*5*). Such ***misinterpretation*** could happen because the system was incapable of recognizing phrases out of its vocabulary. Other common issues were related to ***word spelling*** and ***punctuation*** in text fields, and participants were particularly frustrated by incorrect punctuation: "*spelling errors were kind expected, but the annoying thing was that it kept interpreting my pauses as periods when they should have been commas*" (P8). When spelling and punctuation errors in text fields occurred, participants usually edited them using touch input, although it might not be easy to switch from speech to touch (P8, P9, P13, P17): "*my cursor wasn't very accurate, so even it's just a minor spelling error, it takes so many clicks to get it*" (P8). P9 and P17 felt that switching between input modalities could interrupt the "*flow*" of speaking: "*I think there is a flow when you are in the mode of speaking out all the details about your feelings... and if I chose to speak, I was not ready to type anyway, and it could really interrupt my thoughts*" (P17). Sometimes, participants did not even bother to fix the errors: "*If I wasn't feeling very patient, I would leave them there. It felt like some extra work if you have to go and edit*" (P9).

5.2.4 Data Mismatching Issues. Besides speech recognition issues, NoteWordy could mismatch extracted information to irrelevant text fields when participants were using GS. By examining the utterances right before the use of the clear button , we identified 77 text mismatching instances. Among these instances, 49 (63.6%) were caused by filler words at the beginning or end of the utterance. For example, the utterance "*From 9 to 11 a.m., I was doing work-related tasks at home. Yeah*" included a filler word "*Yeah*," which was unexpectedly placed into task description (PD4). The other 28 (36.4%) instances all occurred in Productivity Diary, where participants intended to capture their productivity score (PD5) by including the word "*feel*" or its other forms (e.g., "*I felt productive*"). While the productivity score was correctly recognized, the text segment also appeared in feelings (PD7). Although participants could quickly clear the wrong text with the clear button, such mismatching issues still cost additional input effort.

5.3 RQ3: Data Richness

By analyzing the data richness of responses to free-form text fields, we grouped each response into one of the three categories: *generality*, *specifics*, and *specifics with additional contexts*. We note that *specifics with additional contexts* were responses that already provided specific details, along with *additional contexts* which can be removed without affecting the completeness of the response [40, 57, 65]. In Break Diary, we found that most of the responses in Break Diary were under the specifics category (BD2: 96.3%; BD3: 91.4%), suggesting a small variation in data richness. Thus, we focused on examining the responses in Productivity Diary.

Table 5 describes how we characterized data richness of each text field with examples, and table 6 summarizes the number of responses in each category and the input modalities involved. *Responses involving speech input* include responses that entered by speech-only input as well as those that entered by both touch and speech input. We did not differentiate these two types of responses because (1) a majority of the responses entered by both modalities were captured by speech input and slightly edited by touch input; (2) the number of these responses only took a small proportion (PD4: 9.3%, PD6: 8.9%, PD7: 8.5%). We also excluded 51 (1.6%) responses in Productivity Diary digressed from the original questions (e.g., answering "*My family felt happy because they had missed me*" to feelings and why (PD7)) from the regression analysis.

Our logistic regression showed that input modality tended to affect the data richness of all three text fields: task description (R^2 = .15, p < .001), productivity rationale (R^2 = .15, p < .001), and feelings (R^2 = .10, p < .001). In task description (PD4) and productivity rationale (PD6), responses involving speech input were more likely to be specific (PD4: OR = 3.79, p < .001; PD6: OR = 2.16, p = .002) and include additional contexts (PD4: OR = 3.0, p < .001; PD6: OR = 4.18, p < .001). In feelings (PD7), although responses involving speech input were not necessarily more likely to be specific (OR = 1.20, p = .36), they tended to include additional contexts (OR = 2.12, p = .03). These

Table 5. The definition and examples of data richness categorization for each text field. In the examples, we highlighted *additional contexts* in blue.

Text field	Generality	Specifics	Specifics with additional contexts
Task description (PD4)	General task type without details "*had a meeting*" "*coding*"	Specific about task activities "*Met the team to discuss mockup design*" "*Writing python code for my class*"	Specific about the task with additional contexts other than time and location asked in the diary "*I attended a UX meeting with other designers. We shared some case studies applying design thinking and talked to the BA team for next steps*"
Productivity rationale (PD6)	Vague about the rationale "*It's not the most* productive time" "*My productivity same as before*"	Rationale clearly explaining why they were productive or not "*Because I got everything done in the time expected without distraction*"	Clearly explained why they were productive or not and elaborated the response with additional contexts (e.g., task procedure, upcoming events) "*This meeting went really well and we had a great discussion. There were no instances of unresolved questions or topics in preparation for our Thursday morning meeting*"
Feelings (PD7)	Vague about why they felt in certain ways "*I felt challenged and frustrated*"	Specific reasons explaining how they felt and why "*I felt great during the task because I was caffeinated enough and I had a good conversation with my student*"	Clearly explained how they felt and why and elaborated the response with additional contexts (e.g., emotion fluctuation, long-term plans) "*I felt discouraged at first because I didn't know what I would write about, then I felt inspired because I found a theme. Then I felt really happy because I was able to submit the assignment. Overall I felt proud for completing a task I had considered skipping and believe I did a good job*"

Table 6. Input modalities (responses entered by touch-only input versus responses involving speech input) x data richness (generality, specificity, and specificity with additional contexts) for each text field in Productivity Diary. Note that we excluded responses that digressed from the questions.

Text field	Responses with touch-only input				Responses involving speech input			
	Total	Generality	Specificity	Specificity with additional contexts	**Total**	Generality	Specificity	Specificity with additional contexts
Task description (PD4)	**479**	202 (42.2%)	205 (42.8%)	72 (15.0%)	**518**	102 (19.7%)	306 (59.1%)	110 (21.2%)
Productivity rationale (PD6)	**614**	151 (24.6%)	394 (64.2%)	69 (11.2%)	**410**	52 (12.7%)	254 (61.9%)	104 (25.4%)
Feelings (PD7)	**588**	198 (33.7%)	352 (59.9%)	38 (6.4%)	**436**	134 (30.7%)	228 (52.3%)	74 (17.0%)
Total	**1681**	551 (32.8%)	951 (56.6%)	179 (10.6%)	**1364**	288 (21.1%)	788 (57.8%)	288 (21.1%)

findings were corroborated during the interviews, as participants recalled that with speech input, they were inclined to enter more details and express their thoughts more freely (P1, P10, P11, P13, P15, P17): "*In a natural way, I definitely put more using speech, because I can just talk, and typing is more time consuming. Like speech is a more free and natural way for me to express my thoughts, I guess especially for productivity (rationale) and how I felt*" (P1).

6 DISCUSSIONS

In this section, we reflect on the lessons learned from designing and deploying NoteWordy. We also discuss opportunities for better integrating touch and speech input to support self-tracking in various contexts.

6.1 Integrating Touch & Speech to Support Capturing Different Types of Data

We found ***speech*** input significantly reduced the time spent on completing the diary entries and helped enhance richness of free-form text. These findings corroborated prior studies on speech-based food journaling, which suggested that speech input was perceived easy to use and could encourage people to elaborate their responses with additional contexts [40]. The richness of speech input is important for collecting self-reported behaviors and assessments—data that are difficult to automatically capture or interpret due to lack of contextual information (e.g., health symptoms [43], mood [52], reflective thoughts [70]). ***Touch*** input was frequently used for capturing structured data including timespan, multiple choice, and Likert scale questions, especially the latter two that require only a single tap. Participants were also more comfortable with touch input in public spaces, where they concerned about privacy. Even for the same data field, their modality preferences might differ depending on the socio-technical context where data capture happened (e.g., not wanting colleagues to hear about non-productive tasks). Our study demonstrated the effectiveness of integrating both touch and speech input to capture multiple types of data in broader scenarios.

On the other hand, we noticed that some participants seldom used speech input despite having a neutral or positive experience with speech interfaces. Besides privacy concerns and recognition issues that echoed with prior research [39–41], our participants talked about the "atypical *(social) norm*" of using speech input to capture personal data. We suspect that this might be due to (1) the design of NoteWordy's data capture interface—an entry form that is typically completed by touch input; and (2) the private nature of personal data that prevented participants from speaking

aloud. As a means of encouraging people to take advantage of speech input, we can design a more interactive data capture interface—for example, introducing a conversational agent (CA) to engage people in "conversations" about their personal data [36]. To account for privacy concerns and individual preferences, more research is needed to investigate how to design CAs with different attributes (e.g., visual, language style) to support capturing various types of personal data.

6.2 Supporting Efficient Multi-Data Capture With Speech Input

Our participants used GS to capture multiple data fields in various ways. They acknowledged that GS was fast and convenient, especially with Break Diary, in which all the data fields were displayed on one screen and can be easily linked together when using natural language. However, it was not always intuitive to include multiple data fields in one sentence (e.g., "*working on a work-related task at workplace*"). In such cases, using GS can be redundant and add extra input burden. This finding suggests the importance for NLI-based data capture tools to arrange semantically-related data fields on one screen, so that people can easily skim what data to capture and then naturally compose their utterances.

Another challenge related to GS adoption was unfamiliarity. Participants were unsure about how to properly phrase an utterance containing multiple data fields or felt mentally taxing to come up with such an utterance. While there was no statistic trend showing how participants' modality preference changed over time, our interview revealed that even those who were able to adopt GS (e.g., P13) still experienced a "learning period" that took about a week. These findings showed that although we provided preemptive guides in NoteWordy (see Figure 5b) and conducted practice sessions during the tutorial, these guides and practices might not be enough for participants to overcome the adoption barriers. As such, people may need more informative "prompts" to get used to GS. One opportune prompting moment is when people are using LS: the system can suggest that the current data field being entered can also be captured together with others via GS with utterance examples.

6.3 Enabling Effective & Flexible Data Editing With Multimodal Input

The data capture burden associated with speech recognition issues posed new challenges in supporting multimodal data editing. For example, when timespan data was partially misrecognized (incorrect start or end time), participants tended to manually update the incorrect time point with touch input, because they did not want to repeat both the start and end time with speech input. When recognition issues occurred in text fields, even a minor spelling error would require extra effort to fix (e.g., place the cursor to the right position, remove the wrong spelling, and type). Such editing solutions are not ideal, because when participants chose to use speech input, oftentimes they were not ready to type or touch the screen.

To support more effective and flexible data editing, we can enable speech-based editing with minimal involvement of touch input [24]. For timespan data, people can select the target time point and update it with speech input (e.g., pressing on the end time field and say "*9:15 p.m.*" or the minute field and say "*15*"), or adopt the command-based approach by specifying the component that they intend to update in the utterances (e.g., say "Updating the *end time* to *9:15 p.m.*"). Similarly, for text fields, people can press on the area near the target words, and specify how they want to update existing text (e.g., say "Correct *master* to *semester*").

6.4 Adapting Speech Recognizers for Various Self-Tracking Activities

Like many speech input systems, NoteWordy embeded a commercial service as a speech-to-text recognizer. However, these speech recognizers are trained with context-agnostic dataset and are not fine-tuned for self-tracking activities. Without considering the personal data capture

context, the speech recognizer that we embedded sometimes failed to handle the ambiguity in word pronunciations. Hence, data fields such as timespan and Likert scale are particularly vulnerable to speech recognition errors: if one of the keywords was misrecognized, the entire utterance could became invalid (e.g., "*7 to 9*" being recognized as "*729*") [30].

Although speech recognition services such as Microsoft Cognitive Speech API [3] and Google Cloud Speech-to-text [1] allow developers to fine-tune a recognizer by uploading their own training dataset, these services often require these dataset to be large enough to cover multiple speech variances (e.g., accents, dialects), which are not yet available in the domain of self-tracking. We call for research efforts to contribute to contextualized speech data from diverse self-tracking activities, including but not limited to date and time [33], commonly used labels of Likert scale (e.g., sleep quality [14], stress level [46]), and units for describing daily activities (e.g., cups of coffee [14], exercise repetitions [41]). These data will serve as valuable training resources, allowing researchers to conduct high quality and reproducible error analysis with less effort [25, 73], so that we can better adapt existing speech recognizers for a broader range of personal data capture.

6.5 Limitations and Future Work

To support multiple data capture via GS, we relied on rules and keywords to process and categorize the text transcribed from speech input, which cannot be generalized to other data capture regimens beyond the Productivity Diary and Break Diary in our study. However, these two diaries equipped with touch, LS, and GS input can be considered as a test bed to situate people and to gather empirical insights into real-world experience in capturing personal data with touch and speech input.

Unlike prior work that designates speech input to capture a single type or unit of data, our study provides a deep understanding of how people use touch and speech input to capture multiple types of data. We also examined how touch and speech input on smartphones affected the data capture burden and data richness. Going forward, we see the opportunities to employ multimodal self-tracking in a domain-agnostic context, where people can customize and update which data to capture in what types. This requires researchers to improve the speech input pipeline leveraging the state-of-the-art NLP techniques to "understand" natural language input that describes various self-tracking activities. Moreover, this work holds the promise to improve the accessibility of personal data collection tools for people with vision or motor impairment by enabling multimodal data capture with touch and speech input.

7 CONCLUSION

We reported a two-week long data collection study with NoteWordy, a multimodal smartphone application integrating touch and speech input to capture different types of data, in the context of productivity tracking. During the study, 17 working graduate students collected data about their work- and school-related tasks and breaks using touch and speech input. With both quantitative and qualitative results, we demonstrated how data types interplaying with participants' personal preferences and social surroundings contributed to the ways that they used touch and speech input. Participants praised the convenience of speech input, especially GS for capturing multiple data fields; they also valued touch input, including manual typing, for capturing long and complicated thoughts and preserving their privacy. In addition, we found that speech input significantly reduced diary entry time and enriched the data in free-form text fields. With the lessons learned, we discuss implications for leveraging the strengths of touch and speech input to better support personal data capture in self-tracking contexts, and opportunities for adapting speech recognizers to capture various self-tracking data. We hope this work can inspire researchers working at the intersections of personal informatics and multimodal interaction to design for low-burden, rich, and engaging data capture experience.

8 ACKNOWLEDGMENT

We thank our participants for their time and interests in this study. We also thank Hernisa Kacorri, Beth St. Jean, and Philip Resnik for their thoughtful feedback. This research was supported by National Science Foundation under Award Number #1753452.

REFERENCES

[1] Google Cloud - Send a recognition request with speech adaptation. https://cloud.google.com/speech-to-text/docs/context-strength. Accessed: 2022-09-30.
[2] Kotlin. https://kotlinlang.org/. Accessed: 2022-09-30.
[3] Microsoft Cognitive Service - Prepare data for Custom Speech. https://docs.microsoft.com/en-us/azure/cognitive-services/speech-service/how-to-custom-speech-test-and-train. Accessed: 2022-09-30.
[4] Microsoft Cognitive Service - Speech to Text. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/. Accessed: 2022-09-30.
[5] Natty. http://natty.joestelmach.com/. Accessed: 2022-09-30.
[6] Smile - Statistical Machine Intelligence and Learning Engine. https://haifengl.github.io/nlp.html. Accessed: 2022-09-30.
[7] Amazon Alexa. https://alexa.amazon.com/. Accessed: 2022-09-30.
[8] Otter.ai. https://otter.ai/. Accessed: 2022-09-30.
[9] Peter Bates and Lori Goff. 2012. The invisible student: Benefits and challenges of part-time doctoral studies. *Alberta Journal of Educational Research* 58, 3 (2012), 368–380. https://doi.org/10.11575/ajer.v58i3.55628
[10] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012). https://doi.org/10.1037/13620-004
[11] Hancheng Cao, Chia-Jung Lee, Shamsi Iqbal, Mary Czerwinski, Priscilla NY Wong, Sean Rintel, Brent Hecht, Jaime Teevan, and Longqi Yang. 2021. Large scale analysis of multitasking behavior during remote meetings. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3411764.3445243
[12] Barbara L Chalfonte, Robert S Fish, and Robert E Kraut. 1991. Expressive richness: a comparison of speech and text as media for revision. In *Proceedings of the 1991 CHI Conference on Human Factors in Computing Systems*. 21–26. https://doi.org/10.1145/108844.108848
[13] Eugene Charniak and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*. https://www.aclweb.org/anthology/N01-1016
[14] 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. https://doi.org/10.1145/2750858.2804266
[15] Rajeev Darolia. 2014. Working (and studying) day and night: Heterogeneous effects of working on the academic performance of full-time and part-time students. *Economics of Education Review* 38 (2014), 38–50. https://doi.org/10.1016/j.econedurev.2013.10.004
[16] Marika De Bruijne and Arnaud Wijnant. 2013. Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computer-assisted web survey. *Social Science Computer Review* 31, 4 (2013), 482–504. https://doi.org/10.1177/0894439313483976
[17] Marion Dunagan. 2012. *Coping strategies of part-time MBA students: The role of boundary management*. University of Arkansas.
[18] Daniel A Epstein, Daniel Avrahami, and Jacob T Biehl. 2016. Taking 5: Work-breaks, productivity, and opportunities for personal informatics for knowledge workers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 673–684. https://doi.org/10.1145/2858036.2858066
[19] Allyson Ettinger, Sudha Rao, Hal Daumé III, and Emily M Bender. 2017. Towards linguistically generalizable NLP systems: A workshop and shared task. *arXiv preprint arXiv:1711.01505* (2017). https://arxiv.org/abs/1711.01505
[20] Asbjørn Følstad and Petter Bae Brandtzæg. 2017. Chatbots and the new world of HCI. *interactions* 24, 4 (2017), 38–42. https://dl.acm.org/doi/10.1145/3085558
[21] Dayne Freitag. 2000. Machine learning for information extraction in informal domains. *Machine learning* 39, 2 (2000), 169–202. https://doi.org/10.1023/A:1007601113994
[22] Genesant Technologies, Inc. Talk-to-Track. https://www.talktotrack.com/ Accessed: 2022-09-30.
[23] VI Gerchikov. 2000. The phenomenon of the working college student. *Russian Education & Society* 42, 6 (2000), 67–84. https://doi.org/10.2753/RES1060-9393420667
[24] Debjyoti Ghosh, Can Liu, Shengdong Zhao, and Kotaro Hara. 2020. Commanding and Re-Dictation: Developing Eyes-Free Voice-Based Interaction for Editing Dictated Text. *ACM Transactions on Computer-Human Interaction (TOCHI)* 27, 4 (2020), 1–31. https://doi.org/10.1145/3390889
[25] Sharon Goldwater, Dan Jurafsky, and Christopher D Manning. 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication* 52, 3 (2010), 181–200.

https://doi.org/10.1016/j.specom.2009.10.001

[26] ID Hill. 1983. Natural language versus computer language. *Designing for human-computer communication* (1983), 55–72.

[27] Alexis Hiniker, Sungsoo Hong, Tadayoshi Kohno, and Julie A Kientz. 2016. Mytime: Designing and evaluating an intervention for smartphone non-use. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 4746–4757. https://doi.org/10.1145/2858036.2858403

[28] Journify, Inc. Journify. https://journify.co/ Accessed: 2022-09-30.

[29] Young-Ho Kim, Eun Kyoung Choe, Bongshin Lee, and Jinwook Seo. 2019. Understanding personal productivity: How knowledge workers define, evaluate, and reflect on their productivity. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3290605.3300845

[30] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E. Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. MyMove: Facilitating Older Adults to Collect In-Situ Activity Labels on a Smartwatch with Speech. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 416, 21 pages. https://doi.org/10.1145/3491102.3517457

[31] Young-Ho Kim, Jae Ho Jeon, Bongshin Lee, Eun Kyoung Choe, and Jinwook Seo. 2017. OmniTrack: A flexible self-tracking approach leveraging semi-automated tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–28. https://doi.org/10.1145/3130930

[32] Young-Ho Kim, Bongshin Lee, Jinwook Seo, and Eun Kyoung Choe. OmniTrack for Research: A Research Platform for Streamlining Mobile-based In Situ Data Collection. https://omnitrack.github.io/research

[33] Young-Ho Kim, Bongshin Lee, Arjun Srinivasan, and Eun Kyoung Choe. 2021. Data@ Hand: Fostering Visual Exploration of Personal Data on Smartphones Leveraging Speech and Touch Interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3411764.3445421

[34] Philipp Kirschthaler, Martin Porcheron, and Joel E Fischer. 2020. What can i say? effects of discoverability in vuis on task performance and user experience. In *Proceedings of the 2nd Conference on Conversational User Interfaces*. 1–9. https://doi.org/10.1145/3405755.3406119

[35] Wolfgang Klein. 2018. The Basic Variety (or: Couldn't natural languages be much simpler?). In *Looking at Language*. De Gruyter Mouton, 403–465. https://doi.org/10.1515/9783110549119-016

[36] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. 2018. Reflection Companion: A Conversational System for Engaging Users in Reflection on Physical Activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (2018), 70. http://doi.org/10.1145/3214273

[37] Mandy Korpusik, Salima Taylor, Sai Krupa Das, Cheryl Gilhooly, Susan Roberts, and James Glass. 2019. A food logging system for iOS with natural spoken language meal descriptions (P21-009-19). *Current developments in nutrition* 3, Supplement_1 (2019), nzz041–P21. https://doi.org/10.1093/cdn/nzz041.P21-009-19

[38] Anuj Kumar, Tim Paek, and Bongshin Lee. 2012. Voice typing: a new speech interaction model for dictation on touchscreen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2277–2286. https://doi.org/10.1145/2207676.2208386

[39] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–31. https://doi.org/10.1145/3274371

[40] Yuhan Luo, Young-Ho Kim, Bongshin Lee, Naeemul Hassan, and Eun Kyoung Choe. 2021. FoodScrap: Promoting Rich Data Capture and Reflective Food Journaling Through Speech Input.. In *Proceedings of the 2021 Conference on Designing Interactive System*. ACM. https://doi.org/10.1145/3461778.3462074

[41] Yuhan Luo, Bongshin Lee, and Eun Kyoung Choe. 2020. TandemTrack: shaping consistent exercise experience by complementing a mobile app with a smart speaker. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–13. https://doi.org/10.1145/3313831.3376616

[42] Yuhan Luo, Bongshin Lee, Donghee Yvette Wohn, Amanda L Rebar, David E Conroy, and Eun Kyoung Choe. 2018. Time for break: Understanding information workers' sedentary behavior through a break prompting system. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14. https://doi.org/10.1145/3173574.3173701

[43] Yuhan Luo, Peiyi Liu, and Eun Kyoung Choe. 2019. Co-Designing food trackers with dietitians: Identifying design opportunities for food tracker customization. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3290605.3300822

[44] Holger Lütters, Malte Friedrich-Freksa, and Marc Egger. 2018. Effects of speech assistance in online questionnaires. In *General Online Research Conference, Cologne, Germany*.

[45] Gloria Mark, Mary Czerwinski, Shamsi Iqbal, and Paul Johns. 2016. Workplace indicators of mood: Behavioral and cognitive correlates of mood among information workers. In *Proceedings of the 6th International Conference on Digital Health Conference*. 29–36. https://doi.org/10.1145/2896338.2896360

[46] Gloria Mark, Daniela Gudith, and Ulrich Klocke. 2008. The cost of interrupted work: more speed and stress. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 107–110. https://doi.org/10.1145/1357054.1357072

[47] Gloria Mark, Shamsi Iqbal, and Mary Czerwinski. 2017. How blocking distractions affects workplace focus and productivity. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 928–934. https://doi.org/10.1145/3123024.3124558

[48] Midnight Plan, Inc. MurMur. https://play.google.com/store/apps/details?id=com.midnightplan.murmur Accessed: 2022-09-30.

[49] Christine Mollen. Trying to Fit In: Barriers to Degree Completion for Part-Time Graduate Students.

[50] Cosmin Munteanu, Matt Jones, Sharon Oviatt, Stephen Brewster, Gerald Penn, Steve Whittaker, Nitendra Rajput, and Amit Nanavati. 2013. We need to talk: HCI and the delicate topic of spoken language interaction. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 2459–2464. https://doi.org/10.1145/2468356.2468803

[51] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R Cowan. [n.d.]. Design guidelines for hands-free speech interaction. In *MobileHCI 2018*. 269–276. https://doi.org/10.1145/3236112.3236149

[52] Elizabeth L Murnane, Dan Cosley, Pamara Chang, Shion Guha, Ellen Frank, Geri Gay, and Mark Matthews. 2016. Self-monitoring practices, attitudes, and needs of individuals with bipolar disorder: implications for the design of technologies to manage mental health. *Journal of the American Medical Informatics Association* 23, 3 (2016), 477–484. https://doi.org/10.1093/jamia/ocv165

[53] Jessica M Nicklin, Emily J Meachon, and Laurel A McNall. 2019. Balancing work, school, and personal life among graduate students: A positive psychology approach. *Applied Research in Quality of Life* 14, 5 (2019), 1265–1286. https://doi.org/10.1007/s11482-018-9650-z

[54] Fatma Őzcan, Abdul Quamar, Jaydeep Sen, Chuan Lei, and Vasilis Efthymiou. 2020. State of the art and open challenges in natural language interfaces to data. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*. 2629–2636. https://doi.org/10.1145/3318464.3383128

[55] Denise Pfeiffer. 2001. *Academic and environmental stress among undergraduate and graduate college students: A literature review*. Ph.D. Dissertation. https://minds.wisconsin.edu/bitstream/handle/1793/40121/2001pfeifferd.pdf?sequence=1

[56] Jakub Piskorski and Roman Yangarber. 2013. Information extraction: Past, present and future. In *Multi-source, multilingual information extraction and summarization*. Springer, 23–49. https://doi.org/10.1007/978-3-642-28569-1_2

[57] Melanie Revilla, Mick P Couper, Oriol J Bosch, and Marc Asensio. 2020. Testing the use of voice input in a smartphone web survey. *Social Science Computer Review* 38, 2 (2020), 207–224. https://doi.org/10.1177/0894439318810715

[58] Verónica Rivera-Pelayo, Angela Fessl, Lars Müller, and Viktoria Pammer. 2017. Introducing mood self-tracking at work: Empirical insights from call centers. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (2017), 1–28. https://doi.org/10.1145/3014058

[59] Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James A Landay. 2018. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–23. https://doi.org/10.1145/3161187

[60] Johan Schalkwyk, Doug Beeferman, Françoise Beaufays, Bill Byrne, Ciprian Chelba, Mike Cohen, Maryam Kamvar, and Brian Strope. 2010. "Your word is my command": Google search by voice: A case study. In *Advances in speech recognition*. Springer, 61–90. https://doi.org/10.1007/978-1-4419-5951-5_4

[61] Michael F Schober, Frederick G Conrad, Christopher Antoun, Patrick Ehlen, Stefanie Fail, Andrew L Hupp, Michael Johnston, Lucas Vickers, H Yanna Yan, and Chan Zhang. 2015. Precision and disclosure in text and voice interviews on smartphones. *PloS one* 10, 6 (2015), e0128337. https://doi.org/10.1371/journal.pone.0128337

[62] Katie Seaborn and Jacqueline Urakami. 2021. Measuring Voice UX Quantitatively: A Rapid Review. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8. https://doi.org/10.1145/3411763.3451712

[63] Holly Seirup and Sage Rose. 2011. Exploring the effects of hope on GPA and retention among college undergraduate students on academic probation. *Education Research International* 2011 (2011). https://doi.org/10.1155/2011/381429

[64] Lucas M Silva and Daniel A Epstein. 2021. Investigating Preferred Food Description Practices in Digital Food Journaling. In *Proceedings of the 2021 Conference on Designing Interactive System*. ACM. https://doi.org/10.1145/3461778.3462145

[65] Jolene D Smyth, Don A Dillman, Leah Melani Christian, and Mallory McBride. 2009. Open-ended questions in web surveys: Can increasing the size of answer boxes and providing extra verbal instructions improve response quality? *Public Opinion Quarterly* 73, 2 (2009), 325–337. https://doi.org/10.1093/poq/nfp029

[66] Arjun Srinivasan, Bongshin Lee, Nathalie Henry Riche, Steven M Drucker, and Ken Hinckley. 2020. InChorus: Designing consistent multimodal interactions for data visualization on tablet devices. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3313831.3376782

[67] Arjun Srinivasan and John Stasko. 2020. How to ask what to say?: Strategies for evaluating natural language interfaces for data visualization. *IEEE Computer Graphics and Applications* 40, 4 (2020), 96–103. https://doi.org/10.1109/MCG.

2020.2986902

[68] TeamViewer. TeamViewer QuickSupport. https://www.teamviewer.com/en-us/info/quicksupport/. Accessed: 2022-09-30.

[69] Helma Torkamaan and Jürgen Ziegler. 2020. Exploring chatbot user interfaces for mood measurement: a study of validity and user experience. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*. 135–138. https://doi.org/10.1145/3410530.3414395

[70] Ciaran B Trace and Yan Zhang. 2021. Minding the gap: Creating meaning from missing and anomalous data. *Information & Culture* 56, 2 (2021), 178–216. https://doi.org/10.7560/IC56204

[71] Aditya Vashistha, Pooja Sethi, and Richard Anderson. 2018. BSpeak: An accessible voice-based crowdsourcing marketplace for low-income blind people. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3173574.3173631

[72] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer. 2017. An analysis of environment, microphone and data simulation mismatches in robust speech recognition. *Computer Speech & Language* 46 (2017), 535–557. https://doi.org/10.1016/j.csl.2016.11.005

[73] Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2019. Errudite: Scalable, reproducible, and testable error analysis. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/P19-1073

[74] Yixuan Zhang and Andrea G Parker. 2020. Eat4Thought: A Design of Food Journaling. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8. https://doi.org/10.1145/3334480.3383044