Investigating the Tradeoffs of Everyday Text-Entry Collection Methods

André Rodrigues
afrodrigues@fc.ul.pt
LASIGE, Faculdade de Ciências, Universidade de Lisboa
Portugal

Hugo Nicolau
hugo.nicolau@tecnico.ulisboa.pt
ITI/LARSyS, Instituto Superior Técnico, Universidade de Lisboa
Portugal

André Santos
abranco@lasige.di.fc.ul.pt
LASIGE, Faculdade de Ciências, Universidade de Lisboa
Portugal

Diogo Branco
djbranco@fc.ul.pt
LASIGE, Faculdade de Ciências, Universidade de Lisboa
Portugal

Jay Rainey
j.rainey2@newcastle.ac.uk
Open Lab, Newcastle University
United Kingdom

David Verweij
david.verweij@newcastle.ac.uk
Open Lab, Newcastle University
United Kingdom

Jan David Smeddinck
jan.smeddinck@newcastle.ac.uk
Open Lab, Newcastle University
United Kingdom

Kyle Montague
kyle.montague@northumbria.ac.uk
Northumbria University
United Kingdom

Tiago Guerreiro
tjvg@di.fc.ul.pt
LASIGE, Faculdade de Ciências, Universidade de Lisboa
Portugal

ABSTRACT

Typing on mobile devices is a common and complex task. The act of typing itself thereby encodes rich information, such as the typing method, the context it is performed in, and individual traits of the person typing. Researchers are increasingly using a selection or combination of experience sampling and passive sensing methods in real-world settings to examine typing behaviours. However, there is limited understanding of the effects these methods have on measures of input speed, typing behaviours, compliance, perceived trust and privacy. In this paper, we investigate the tradeoffs of everyday data collection methods. We contribute empirical results from a four-week field study (N=26). Here, participants contributed by transcribing, composing, passively having sentences analyzed and reflecting on their contributions. We present a tradeoff analysis of these data collection methods, discuss their impact on text-entry applications, and contribute a flexible research platform for in the wild text-entry studies.

CCS CONCEPTS

• **Human-centered computing** → **Smartphones**; **Field studies**; **Text input**.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3501908

KEYWORDS

text-entry, in-the-wild, trade-offs, touch behaviours, user experience, performance, data collection

ACM Reference Format:
André Rodrigues, Hugo Nicolau, André Santos, Diogo Branco, Jay Rainey, David Verweij, Jan David Smeddinck, Kyle Montague, and Tiago Guerreiro. 2022. Investigating the Tradeoffs of Everyday Text-Entry Collection Methods. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3491102.3501908

1 INTRODUCTION

The widespread adoption of mobile devices and the trend towards asynchronous means of communication (e.g., SMS, email, social media) has led to a massive increase in the amount of typing data we generate in our everyday lives. Thus, in the past decade, we have witnessed a growing interest in leveraging rich data from everyday typing behaviour for multiple research purposes. For example, researchers aim to understand real-world typing behaviours to *improve performance* of keyboards in terms of speed and accuracy [9, 21]. Others have also started collecting behavioural data for personalised and *context-aware keyboards* that adapt to individual typing patterns and situations [24, 25]. More recently, in the field of *biometrics*, research shows that everyday typing activities can be used as authentication methods due to individual differences between users [13, 19, 30]. In the field of *linguistics*, text-entry data is often leveraged to study the impact of computer-mediated communication on language and variations across ethnographies [8, 56]. Text input analysis can also be used as a *health monitoring* tool, showing potential for early disease detection [5, 18] and to assess stress [16], fatigue [3], and inebriation [41].

While text input analysis offers opportunities in multiple research domains collecting everyday typing data remains a key challenge. There are two distinct approaches to study real-world typing

behaviours: (1) *experience sampling*, where users are prompted to interact with an application or website throughout the day to perform text-entry tasks (i.e., transcription or compositions); and (2) *passive sensing*, where background services or system-wide tools (e.g., keyboards) are installed on users' devices to collect natural behaviours from everyday typing tasks. While both approaches have their merit, neither is an ideal solution. Collecting users typing data raises several challenges and concerns with respect to data privacy, user effort, trust, compliance, and data accuracy.

Prior research has explored a multitude of methods to collect in the wild typing data. Notable approaches include the use of experience sampling methods [13], leveraging always-available online platforms [17, 46], to more embedded solutions that collect data passively [23, 43]. When choosing between data collection methods, researchers are forced to consider the complex tradeoff between restricting data collection to respect users' privacy; and collecting comprehensive, unbiased, and natural text-entry data. Despite being a topic of great interest to the HCI community, we have limited understanding of the effects of these data collection methods on the resulting measures of users' input speed, typing behaviours, compliance, perceived trust and privacy.

In this paper, we investigate the tradeoffs of everyday data collection methods across multiple levels of analysis: text-entry performance, touchscreen behaviours, and perceived user experience. As more research and applications are leveraging everyday text-entry data, it is crucial to study both the attitudes and perceptions of end users towards privacy, trust and compliance, while providing effective data collection methods that can accurately capture natural input behaviours and performance. We thus aim to answer the overarching research question: what are the tradeoffs between experience sampling and passive sensing in relation to 1) compliance and coverage, 2) user's typing performance, and 3) the user experience?

To answer this question, we conducted a month-long user study comprised of two remote researcher observed sessions (i.e., briefing and debriefing) and four weeks of everyday text-entry data. During this period, we examined three data collection methods: first, *experience sampling with transcription tasks* (ExpT), which allow accurate and rigorous quantification of speed and errors. Second, *experience sampling with composition tasks* (ExpC) (e.g, "write a text message describing the activities you performed throughout the day"), which captured the cognitive aspects related to the process of text generation; however, it introduced uncertainty in computing error rates. Finally, *passive sensing* (PasS) that aimed to collect unbiased everyday typing data but introduced additional issues related to participants' privacy and trust. Finally, we offer an Android-based research platform for in the wild text-entry studies that supports multiple data collection methods while ensuring user privacy.

Results show that there are multiple significant differences in typing performance and behaviour metrics (cf. section 5.2), e.g. concerning speed, error rates, and touch dynamics, as well as notable differences in user perceptions and user experience (cf. section 5.3) between the data collection methods concerning perceived privacy, trust, and acceptance.

The key contributions of this paper are: (1) empirical results into everyday text input collected in a four week field study (N=26); (2) tradeoff analysis of three text-entry data collection methods across multiple levels of analysis (text-entry performance, touch behaviours, and user experience); (3) implications that contextualise data collection methods on text-entry applications; (4) a research platform that supports various data collection methods for in the wild text-entry studies; and (5) an everyday text input dataset consisting of collected from a four-week field study (N=26). These contributions are relevant to researchers in HCI and in other domains that aim to leverage typing data, as well as more generally to designers and developers of applications that rely on – or facilitate – typing dynamics features. Reported results provide the basis for understanding the tradeoffs between data collection methods, inform the design of future studies, and foster novel approaches for collecting data in-situ.

2 RELATED WORK

We discuss related work along three key topics: first, we analyse why studying everyday typing is important in multiple fields of research. Second, we discuss common data collection methods used in the literature. Finally, we present prior research aiming to understand typing beyond controlled laboratory studies.

2.1 Research Interests in Typing Data

Studying people's everyday typing behaviours is a topic of interest to many fields of research. For example, HCI is often interested in collecting typing data to improve existing keyboards in terms of input accuracy and speed. Prior research aimed at optimising keyboard layouts and letter arrangements based on different cost functions and constraints [9, 10, 44, 63]. Others proposed improving typing performance by adapting to users' typing patterns, emotional state, and context [22, 24, 25, 29, 62]. These adaptive text-entry methods often require large amounts of typing data to build statistical language models that can auto-correct the previous input, complete ongoing typing or predict the next intended word. The adaptations can take many forms, including resizing key targets [27, 29], offsetting touch points [31], creating personalised touch models [59, 62], and providing word suggestions [4]. More recently, input researchers are increasingly interested in assessing the performance of novel input methods outside lab settings as everyday use may have an impact on the typing performance, behaviours, and overall user experience [13, 23, 43, 46, 49].

Typing biometrics is another field interested in understanding and quantifying typing behaviours. Previous research has used individual typing patterns to enhance security and implicitly authenticate users [13, 19, 30, 35]. These techniques generally use touch-specific behavioural features such as touch down and touch up locations alongside timing information. Thus, even if an attacker knows the users' password, an additional implicit security layer can verify the identity based on keystroke dynamics [12, 14]. It is then of utmost importance that typing biometrics are a reflection of in the wild typing behaviours. Building appropriate data collection tools while having a thorough understanding of how collection methods affect typing becomes a key challenge [54].

Studying people's language use in computer-mediated communication also involves collecting typing data. Such data has been used

for multiple purposes, including understanding the impact of technology on language use [7], how language varies across different groups of people [32, 51], how language use changes with age [50], and predicting demographic characteristics and personality from everyday typing data on mobile devices [53] and social networks [26, 61].

Furthermore, data generated from typing tasks has great potential to be used in health research. For instance, keystroke behaviours and linguistic features have been used as a marker with Parkinson's disease [5, 18, 34] as well as multiple sclerosis [37], which shows potential for early disease detection and intervention. Typing data has also been used to assess various health states including stress [16], fatigue [2, 3, 55], and inebriation [41].

Overall, related work shows wide potential for leveraging text-entry data and research in multiple domains. Moreover, researchers are increasingly interested in understanding "true" typing behaviours and moving towards data collection in the wild. Also, different applications may have different requirements for data collection. For example, biometrics and health applications may need to collect low-level touch features of individual keystrokes while input researchers may only be interested in aggregated speed and accuracy measures to assess the performance of input methods. Our research platform supports these studies and can be extended to support bespoke metrics. However, it is unclear, what the most appropriate method to collect everyday typing is. This significant gap in the literature strongly motivated our work.

2.2 Data Collection Methods for Text-Entry

Research on text input has traditionally required controlled laboratory studies for accurate quantification of speed and errors [52]. Users transcribe presented memorable sentences as "quickly and accurately as possible", ensuring that participants only need to copy text, guaranteeing experimental control and reproducibility. Participants are often allowed to correct errors during the typing task by using the backspace, moving the cursor, and using auto-correction features [60, 64]. This protocol allows for accurate computation of input metrics such as words per minute and various error rates. Others have used composition tasks rather than transcription tasks in laboratory settings [20, 57]. Although composition tasks can capture certain cognitive and linguistic aspects related to the process of text generation, computing error rates is a major challenge as researchers must know user intentions to detect deviations [23, 57].

In the last two decades, there has been a growing interest to understand users' everyday computer use in general [6, 11, 15, 45] and particularly for typing [23, 31, 49]. There have been multiple approaches to collecting typing data outside the laboratory. One of the most common methods is to explicitly prompt users throughout the day (i.e. experience sampling) to perform transcription tasks [49]. Researchers can collect input metrics as well as subjective ratings (e.g., level of fatigue). However, such experience sampling requires additional effort from the user, and it does not collect natural typing behaviours nor provide implicit assessments. Overall, one can expect less compliance as more data is requested from users. On the other hand, it benefits from a well-defined structured task, which may provide data with less noise and enable types of analysis that are otherwise not possible.

Others have used similar approaches to experience sampling by embedding transcription tasks in mobile games [31], or large-scale online experiments [46, 48]. However, previous studies outside the laboratory have almost exclusively used transcription tasks. Going beyond transcription tasks in everyday typing is difficult as we need to know the intended text to calculate error metrics. Approaches to estimate intent have been previously proposed and include using dictionaries, search engines, and crowdworkers [23, 43, 57]. These estimates are then used to compute traditional error rate metrics.

An alternative to experience sampling is passive sensing, i.e. collecting natural free text composition from everyday typing activities [23, 36, 43]. While it requires less effort from users, this method can face concerns related to privacy, and overall adherence to the study [35]. Iakovakis et al. [34] restricted the data collected in their work to metrics not related to content such as flight times and hold times. The approach ensures user privacy at the cost of limiting the type of metrics one can extract from typing sessions. Similar solutions include using filtering methods that omit the vast majority of characters from data collection [13] or text abstraction methods that only collect word counts or word categories [8]. While these methods of passive sensing can be useful for specific applications, they do not provide a common data collection framework to study everyday typing holistically. For instance, it would not be feasible to analyse speed-error input tradeoffs with existing tools.

In summary, related work shows an opportunity to devise novel instruments to study everyday typing and advance input research. Existing tools are often limited in collection methods and metrics, particularly when considering analysing unconstrained free text (i.e. natural typing). We contribute a customisable research platform that supports both experience sampling and passive sensing. Both methods can be used to collect speed, error rates, and touch behaviours while always preserving the users' privacy.

2.3 Text-Entry in the Wild

Research has shown that performance, dynamics, and user experience of typing vary between laboratory and real-world settings [13, 23, 43, 49]. Reyal et al. [49] conducted a laboratory and an experience sampling experiment with two mobile keyboards and suggest that both methodologies should be used in conjunction as they can be informative of different aspects of the keyboard experience. Evans and Wobbrock [23] used passive sensing on desktop computers and showed that everyday typing speed is slightly faster and more erroneous than laboratory assessments. Results from Nicolau et al. [43] show the same trend for the everyday typing performance of mobile screen reader users. Zhang et al. [65] conducted laboratorial (between strangers) and longitudinal (replacing the user keyboard) studies to understand the impact of emoji suggestions with different insights from each.

These findings suggest that laboratory studies may not be representative of everyday typing; thus, there has been an increasing number of studies conducted outside the laboratory in recent years. Komninos et al. [36] investigated how mobile input errors emerge in real-world situations and how users deal with them. Results show that errors are common, despite participants using spellcheckers. Contrary to general belief, finger slippage is not a major source of errors. Palin et al. [46] collected data from 37,370 participants

via a web-based platform and demonstrate that auto-correct users tend to be faster while those that rely on predictions tend to be slower. Zhang et al.[66] conducted a sequence of studies going from controlled lab experiments to assessments in the wild to explore phrase level input. Henze et al. [31] show that mobile touch distributions have a systematic skew in relation to the intended key, which can be compensated to improve typing performance. Buschek et al. [13] investigated typing biometrics and suggest that individual differences of typing behaviour show in different typing features in the wild compared to transcription in the lab.

Overall, there is a growing effort to "break away from the lab" for input research. The field has adopted a multitude of data collection methods while recognising the need to understand their effects in real-world typing performance and behaviours [13, 49]. To the best of our knowledge, we contribute the first comparative analysis between data collection methods for everyday touch typing and provide novel insights to input performance, touch behaviours, and user experience (e.g., compliance, perceived trust and privacy).

3 WILDKEY: RESEARCH PLATFORM FOR COLLECTING EVERYDAY TYPING DATA

Wildkey (Figure 1) is a toolkit for conducting in the wild text-entry studies. Researchers and developers can deploy their standalone ecosystems and customise the experience sampling and passive sensing to their study requirements. The toolkit is composed of an *Android Keyboard, a Study Management application (React), and a NoSQL Database (Firebase).* Wildkey is open source[1]. The Wildkey keyboard extends the Android Open Source Project (AOSP) Keyboard [1], thus featuring support for 26 languages, auto-correct, suggestions, and many other typical keyboard customisations. In addition to the standard features, the keyboard is capable of passive sensing and prompting a variety of experience sampling tasks.

Smartphones are becoming an extension of oneself [47] and data privacy is of utmost importance. Approaches that seek to collect text-entry behaviours can have difficulties recruiting and possibly face additional compliance issues. To tackle these challenges, when analysing unconstrained free text, during passive sensing, no raw text is ever stored on-device or in the cloud, nor any data that would allow its reconstruction. All analyses are done on the device, per text-entry trial, and only processed data is stored. To ensure user control over data collection, Wildkey also has a permanently available button on the top left of the keyboard to activate a private mode (i.e., "incognito mode"), which resets every time the keyboard is closed. When active, no data is analysed and no data is stored about usage of the function.

3.1 Data Collection

Wildkey can create four types of experience sampling prompts: transcriptions, compositions, questionnaires, and custom-made tasks. The creation, deployment (to target registered users), and scheduling of all tasks is made through the *Study Management* app. We designed the toolkit to be flexible and support the creation of *Custom-Made* tasks that may be relevant for a specific study protocol (e.g. the Alternate Finger Tapping Test [38]); none were used in this work. Lastly, the keyboard is capable of passive sensing by analysing all text the user is writing regardless of the application. All data collected from Wildkey is synchronised via encrypted transport and stored in a cloud database in a JSON format.

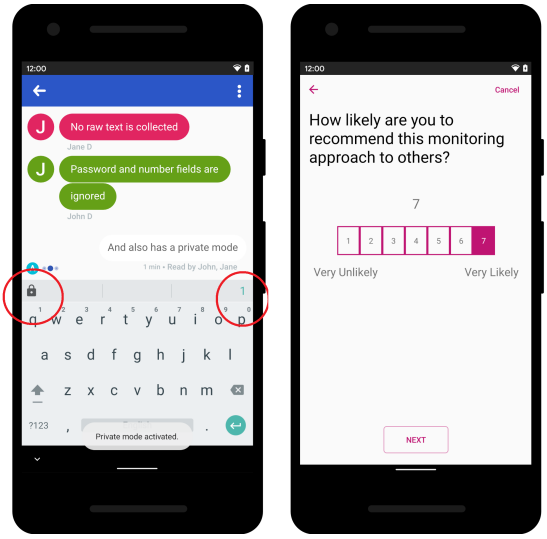

Figure 1: The Wildkey Keyboard with the private mode and the number of pending tasks highlighted. An example of a questionnaire prompted in Wildkey.

3.1.1 Experience Sampling. Users receive a notification when tasks are available to be completed. Additionally, on the keyboard top right corner, a small number appears, indicating a pending task. Users can always open the Wildkey app and check when the next task is scheduled. Depending on how the study was defined, users will have a time interval to complete the task.

Text-entry tasks. Transcriptions and compositions have a similar layout. Sentences/questions are shown at the top of the screen with a text edit field below to transcribe/answer. Users are free to correct, use suggestions and autocomplete. Upon completing a task, the users move to the next trial using the keyboard next/submit/enter button. Time limits can be defined for tasks, as well as character thresholds for composition tasks. In Composition tasks, we estimate the user target sentence. We relied on a similar approach presented in [23] where the intent is calculated using the AOSP Keyboard dictionary and spell checker. Word by word, if it exists, we assume it was the user intent. When the word does not exist, we use the spell checker top recommended suggestions to predict the intended word. The top recommendation is chosen as the intended word to calculate all error rate metrics. We consider all edits to be corrections. In experience sampling, we analyse raw text content and all touch points to compute various performance, touch dynamics, and behaviours (described below) per trial.

Questionnaires. In the wild studies often require users to complete questionnaires, scales, and diaries. To facilitate collecting and cross-referring the data, Wildkey is capable of prompting users to complete customisable questionnaires. Currently, Wildkey supports

[1]under the license Attribution 4.0 International (CC BY 4.0) and available at https://techandpeople.github.io/wildkey/

Slider and Button Scales (Figure 1), Multiple/Single Choice, and Open Questions, all created in the *Study Management* app.

3.1.2 Passive Sensing. During passive sensing, we consider a text-entry trial to start when the user types the first letter after opening the keyboard and ends once the keyboard is closed. The trial is then analysed and resulting metrics are stored. Wildkey does not store touch points, raw text content, app in use, or analyses passwords fields to preserve user privacy. When the user opens the keyboard in a text edit field with text already present, and at some point during the trial moves the cursor to it, we ignore any further changes by marking it as discarded. Text written before each trial is inaccessible to Wildkey; thus, we cannot compute any metrics (without compromising its reliability).

Similar to composition tasks, we estimate the user's intended sentence following the same protocol during passive sensing. Target phrases are not stored and are only used to calculate performance metrics after each trial.

3.2 Study Management

The study management app allows researchers and developers to quickly create a study protocol and deploy it to registered participants. A protocol is created by associating it with a set of tasks and respective sampling schedule. To create transcription/composition tasks, researchers set the phrases/questions to be part of the dataset and define the (time/length) thresholds. For questionnaires, one must first create the individual questions and then group them in a single questionnaire, which can be associated with the protocol.

The study dashboard provides a quick overview of the participants' data and overall performance (e.g., number of characters written, average words per minute), which can be used to track the study progress.

3.3 Touch Entry Metrics

When the user is typing, Wildkey creates 1) an input stream buffer with all keys entered, 2) an array of all actions performed, 3) an array of cursor changes, and 4) a list array of all suggestions given at each key entered. The array of actions consists of corrections (i.e., deletes and substitutions) and entry actions [64]. The array with cursor changes is used to adjust the input stream at the end of the text-entry trial to account for nonsequential changes to the text. Lastly, the suggestion list array is used when predicting target intent in compositions and passive sensing. The input stream is then processed locally by Wildkey to compute all the metrics described below; all other information is discarded. The calculated metrics are synchronised to the cloud database when an internet connection is available.

We calculate speed, error rates, touch dynamics, and text-entry behaviours related metrics for all text-entry trials. For **Speed**, we calculate Words per Minute [39]. For **Error rate** metrics [43, 60], we calculate total, corrected and uncorrected error rates, which are an approximation, and characterisation of the errors users made while writing. For **Typing Dynamics**, Wildkey collects flight and hold time[5], touch offsets [28] and only in experience sampling tasks raw touchpoints . To better characterise users' text-entry behaviours, we collect a variety of **action and character counts** (e.g., selected suggestions, cursor changes, autocorrect). Lastly, Wildkey collects current the keyboard language and, in experience sampling tasks only, raw text.

4 INVESTIGATING THE TRADEOFFS OF TEXT-ENTRY COLLECTION METHODS

Although there is a growing interest in the research community to explore typing data, we have yet to understand the tradeoffs between different text-entry data collection methods in the wild. Our ultimate goal is to contribute with a characterization of different collection methods understanding coverage and compliance, performance, and user experience. In regards to user experience, we were particularly interested in understanding the tradeoffs around privacy, obtrusiveness, and effort. To achieve this, we conducted a four-week longitudinal study with 26 participants, where they were exposed to two text-entry experience sampling methods (i.e. ExpT and ExpC) and a privacy-aware passive sensing (PasS) data collection.

4.1 Participants

Participants were required to be 18 or older, Android users (version 8 or above), and to use at least one communication or social media app (e.g., WhatsApp, Messenger, Facebook) as defined within our IRB approved recruitment criteria. We recruited 26 participants, 15 identified as female and 11 as male. We recruited through social networks, student body mailing lists and relied on early participants to recruit others (snowball sampling). Participants' age ranged from 18 to 63 (M=29.2, SD=13.5) years old. Three participants were swipe typing users, 22 used suggestions when typing, 20 participants typed 10+ times a day, with five reported typing 2-10 times daily, and lastly, one participant who typed 1-10 times a week. We compensated participants for their time with a 20€ gift certificate. The study was conducted in Portugal.

4.2 Procedure

The study comprised two remote observed sessions (briefing and debriefing) with four weeks of free-living text-entry data collection in between (Figure 2). We explored three different data collection methods during these four weeks: 1) experience sampling with transcription tasks (ExpT), 2) experience sampling with composition tasks (ExpC), and 3) passive sensing (PasS). Data from passive sensing (i.e., implicit data) was collected throughout the four weeks. The two middle weeks were counterbalanced between transcription tasks and composition tasks. During each weekend, participants were prompted through Wildkey to complete two questionnaires, one about privacy and one about their acceptance of the method used in the past week. Every Friday, study participants received a personal research contribution email that detailed the number of "characters contributed" to the study that week.

4.2.1 Apparatus. Participants were required to install our experimental Wildkey Android keyboard on their device and use it as their primary keyboard. Wildkey was available on the Google PlayStore and participants were given access to the closed testing branch of the application. We relied on a Firebase realtime database for our cloud storage for Wildkey. Participants could customise their keyboard experience with common features (e.g., vibrate on tap,

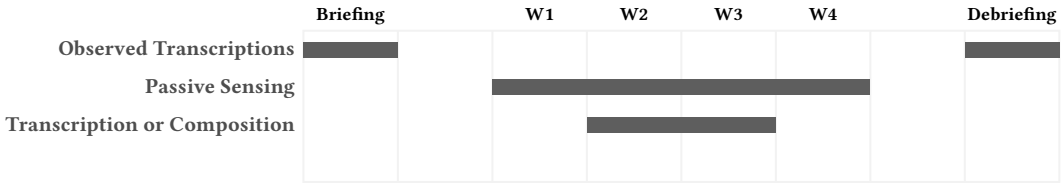

Figure 2: Study timeline overview.

sound on tap), swipe typing was not available as it is not part of the AOSP keyboard package. We relied on the AOSP Portuguese corpus which had 218,470 words for spell checking and word suggestions at the time.

For the two remote sessions, we communicated via Zoom or Discord, depending on participants' preference. Interviews were recorded using Zoom or OBS[2], respectively.

4.2.2 Briefing Session. Prior to the briefing session, participants received an information sheet and a report with mock-up of the type of data collected during the study (available as supplementary material). The documents served to provide context, motivate users, and be transparent about the data collected. Participants were also asked to complete a demographics questionnaire and an informed consent. We scheduled a 30 minutes remote call with each participant. The session started with a brief introduction to the study goals and an overview of the protocol, followed by a remote observed transcription session.

Observed Transcriptions. We asked participants to open the keyboard on a text box and write a short sentence of their choosing. Next, we asked them to open the Wildkey application and start the transcription task that had been scheduled. Participants were asked to transcribe ten sentences (text-entry trial). This data collection step, although not the target of our research, aimed to provide a semi-controlled (i.e. through a remote synchronous call) snapshot of the users' typing performance and behaviours.

Each text-entry trial contained one sentence comprising five words, each word with an average size of 5 characters. We randomly selected ten sentences from a written language corpus of 451 sentences constructed following the methodology from MacKenzie et al. [40]. The character frequency in the corpus had a minimum correlation with the language of 0.97. Participants were encouraged to type as accurately and quickly as possible. Between each trial, participants were free to take short breaks.

We relied on an extension of the unconstrained text-entry protocol [60], where participants are free to correct any errors they encountered, in addition to use autocomplete and word completion suggestions.

4.2.3 Passive Sensing. During the four weeks, Wildkey processed all text-entry trials and calculated metrics locally on the device, calculating error rates, words per minute, flight, hold times, and others.

[2]Open Broadcaster Software (OBS) Studio, available at: https://obsproject.com/, last viewed 2021-09-08

4.2.4 Experience Sampling. For two weeks, with exception of the weekends, participants were prompted to complete text-entry tasks three times a day.

Transcription. During one week, Wildkey prompted participants to complete transcription tasks. They received a notification at the start of the period (9am-1pm, 1pm-7pm, 7pm-11pm) and had the time window to complete the task. They could open the task by pressing the notification, opening the app and seeing pending tasks, or by tapping the icon on the top right of the keyboard that always displayed the number of pending tasks. The task consisted of transcribing three sentences following the protocol described in the Observed Transcription Sessions for text-entry trials.

Composition. Composition followed the same notification and schedule protocol as described in the previous section. In Composition tasks, participants answered three open questions or until they wrote 75 characters or more. We chose a threshold of 75 characters as this matches the amount of data in the transcription condition (3 sentences * 5 words * 5 characters words). Questions were designed to relate to participants daily activities (e.g., "What did you have for lunch?") or tastes (e.g., "How do you like to spend your free time?"). All participants received the same questions in the same order throughout the week.

4.2.5 Acceptance & Privacy Questionnaires. Each weekend, participants were asked to complete a questionnaire about their privacy concerns and acceptance of the data collection method. For PasS the questionnaire was deployed twice, in the first and fourth week, averages were calculated for comparisons. The privacy questions asked to rate how much participants agreed with a set of statements (e.g., "My privacy was protected with this week method", "I was concerned with my privacy this week"). The acceptance questionnaire was based on the Treatment Acceptability Questionnaire [33], which was adapted to refer to monitoring instead of treatment. The questionnaires are available as supplementary material.

4.2.6 Debriefing Session. Before the debriefing session, participants received an email to fill in an online questionnaire alongside a participation report. We designed the questionnaire to encourage participants to reflect on effort, privacy, and preference for the three different collection methods and to rate them post-study. The questionnaire consisted of a set of Likert items and multiple-choice questions (available as supplementary material). The study report was generated from the participants' typing data and contained a preliminary analysis of typing performance (example available as supplementary material). The session consisted of a remote transcription session and a semi-structured interview.

Interview The semi-structured interview lasted on average 20 minutes. We started by asking participants to share how it was to participate in the study. We focused on understanding the participants' perceptions of privacy, effort, and obtrusiveness of each method. We prompted them to reflect on the periodic and debriefing reports as well as their behaviours while participating in the study. Lastly, we discussed participants' confidence and willingness to use the assessment of digital behaviours. Finally, we presented a scenario where text-entry behaviours were monitored for clinical purposes, namely to explore the interplay between privacy concerns, effort, and usefulness in a grounded example.

4.3 Measures

Text-entry performance was measured by analysing the trial's input stream [60]. For passive sensing, we segmented text-entry trials as opening, writing, and closing the keyboard. For transcription and composition tasks, each transcription/answer was considered to be a trial. We report on words per minute (WPM), total, uncorrected and corrected error rates. For **touch dynamics**, we calculate the average Flight Time (i.e., time between releasing a key and taping the next one) and Hold Time (i.e., time between pressing and releasing a key) of a session. Additionally, we report on the **text-entry behaviours** of total use of suggestions and auto-correct. We breakdown each questionnaire to its individual questions and and compare between the three methods. We calculated **compliance** for the experienced sample methods since participants were required to perform 15 tasks for each condition. For passive sensing, we resort to descriptive statistics to characterise the data collected regarding coverage/compliance.

4.4 Data Selection

From a total of 76,235 text entry trials collected in passive sensing, we discarded all trials with less than ten characters (21,978; ∼29%) as they result in inaccurate input measures. We further discarded any sessions in which the average hold time was above one second, or WPM 0, as it indicates an erroneous typing data (197; ∼0.3%). We discarded an additional 15,529 trials (∼20%) as the version of Wildkey used discarded any mismatch between the size in words of the calculated sentence and the phrase on the edit box, this was caused by: multiple cursor changes in passive sensing; text present when opening a edit box and with emojis or other special characters (e.g. unicode). The analyses below focus on the 38,531 renaming trials (∼71% of the trials with ten or more characters). Of the 950 transcription trials, and 417 composition trials collected and used to calculate compliance, we discarded 308 and 80 respectively for erroneous average hold time. The dataset is publicly available[3].

4.5 Design & Analysis

We used a mixed-effects model analysis of variance [42]. Mixed-effect models allow for unbalanced data such as ours. We collected four weeks of passive text-entry data where users typed impromptu and two weeks of experience sampling (where participants were not always compliant). Therefore, we have a different number of trials per participant, per data collection method, and per period of the day. We define a period as the time slot available to answer the experience sampling question (i.e., morning, afternoon and evening). Each trial was attributed one of the defined periods. We modelled the collection method (i.e., Passive | Transcription | Composition | Observed Transcription) and period as fixed effects. In addition, Participant was added as a random effect. In the following sections, we will use PasS, ExpT, ExpC, and ObsT to abbreviate each condition and will refer to passive data as implicit interchangeably. Following the comparisons of the mixed model, we conducted pairwise comparisons with the appropriate Bonferroni corrections to account for the multiple-comparisons. We report on the **estimated marginal means**, which take into account the underlying model of our data; additionally we present a table with all **observed means.**

For the debriefing questionnaire analysis, we relied on the Friedman test, calculate effect size with Kendall's Coefficient of Concordance and applied a post-hoc Wilcoxon test, verifying significance at 0.017 (i.e., 0.05 / 3 methods) to account for multiple tests.

After transcribing all interviews, we conducted a primarily deductive analysis focusing on our concepts of interest around compliance, performance, privacy and effort. The initial codebook was enriched after additional data exploration inductively to include concepts such as: Transparency and Learnability. In addition to conducting interviews and reviewing participants' answers to weekly questionnaires, two researchers familiarised themselves with the data by listening to/reading the transcripts. We created an initial set of codes, deductively informed by our familiarity with the data, and enriched with the concepts that stem from our research interests around privacy, acceptability, and the trade-offs between the methods. The authors discussed the codebook, revised it, and iterated based on the data exploration previously described. Next, two authors independently coded three interviews with the revised codebook, which led to further refinement of coding descriptions, reaching a final Cohen's Kappa agreement of k=0.81 (SD=0.22). The author responsible for conducting most of the interviews proceeded to code the remaining transcripts while simultaneously creating, iterating, and merging identified themes pertaining to our concepts of interest. All authors then used multiple sessions to discuss the themes, associated codes, and identified relationships between themes, which led to findings in the upcoming section Perceived Privacy, Trust, and Acceptance. The codebook is available as supplementary material.

4.5.1 Validating Intent Calculation. To validate our intent calculation algorithm, we compared the computed intended sentences with the existing ground truth, i.e. required sentences from the remote observation and experience sampling transcription tasks. These were the only two instances where we know the true typing intent of users. In summary, we ran the algorithm for all transcribed sentences (remote and experience sampling) and compared error rate metrics between the computed intent and the original required sentence. We found a significant difference in uncorrected error rate (z=-4.29 p<0.001), with estimate M=0.11% (SD=0.21) and ground truth M=1.8% and no significant different in corrected error rate (z=0.0 p=1.00) where the estimation and value were the same. Our algorithm considers as correct any word correctly written which accounts for the (1.7%) difference. One example is someone transcribing the sentence *"It ain't over till the fat lady sings"* as *"It **isn't***

[3]https://github.com/AndreFPRodrigues/Text-Entry-Dataset-

*over **un**til the fat lady sings"*, where there are clear differences between the intended and transcribed sentences, but the transcribed output is a fully correct computed intent. This small and rational difference suggests that our algorithm can effectively estimate intended sentences.

5 RESULTS

We start by providing an overview of the dataset and participants' compliance with data collection methods; then, we report on typing performance, input behaviours, perceived privacy, trust, and acceptance.

5.1 Dataset Overview

After data selection, the 26 participants entered a total of ~1.1 million characters (min(p1)=3940 and max(p7)=245,196) writing over the course of 4-weeks. The dataset contains a total of 38,531 independent text-entry trials in passive sensing, 641 transcription trials, and 337 composition trials. For passive sensing, participants contributed on average ~80 trials per day (SD=92.8, min_p1=4.0, max_p7=462), with an average length of 27.5 characters (SD=20.2 min_p22=18.58, max_p3=47.1). We collected 29,201 suggestion uses, 30,668 auto-corrects, 17,872,89 actions, with 484,407 correction actions and 1,748,637 entry actions.

5.1.1 Compliance with Experience Sampling. Wildkey asked participants to perform 15 composition tasks in a week and 15 transcription tasks in another. Overall participants completed 71.5% (SD=0.21, min=53%) of the composition tasks, and 88.7% (SD= 0.16, min=53%) of the transcription tasks. A paired-samples t-test revealed a significant difference t(25)=4.05, p<.001 between data collection methods on compliance rates. Participants were significantly more compliant with transcriptions than with composition tasks.

5.2 Typing Performance & Behaviours

In this section, we characterise participants text-entry performance across the different methods. We compare input speed, accuracy, touch dynamics, and text-entry behaviours.

5.2.1 Speed. Participants showed heterogeneous typing performance, with our fastest participant (p15) averaging 74.6 WPM (SD=29.0) while p1 averaged 9.5 WPM (SD=4.0). Analysing differences between data collection methods, we found a significant effect on WPM [F3,39769=33.01, p<.001]. A pairwise comparison revealed a significant difference between all methods (p<0.05). The estimate marginal mean M=41.5 WPM (SD=2.9) in PasS, M=47.8 WPM (SD=3.2) in ExpT, M=36.9 WPM (SD=3.4) in ExpC, and M=53.3 WPM (SD=3.2) in ObsT which is to be expected (Table 1 - observed means and standard deviations). The Estimates of Fixed Effects reveal that in comparison with ObsT, PasS is slower 0.9 to 21.3 and ExpC 9.3 to 18.7 times. While in ExpT the effect is not significant it varies from less 20.4 to more 4.4. Participants reported that transcriptions were easy to do *"While transcribing was the easiest, it didn't take any effort" P8*, while composition tasks required participants to reflect and think about what they had to write, requiring cognitive effort. As for ObsT, participants were solely focused on the task at hand, which is not guaranteed in any other data collection method. Lastly, we discarded implicit data with less than ten characters which will most likely be the fastest typing sessions participants have (i.e. quick answers or reply which do not take much cognitive effort). We did not find a significant effect of day period on WPM (p=0.41).

5.2.2 Error Rates. We found a significant effect of Method [$F_{3,39773}$=37.03, p<.001] and interaction of Method*Period [$F_{5,39774}$=3.33, p<.01] on Total Error Rate; of Method on Corrected Error Rate [$F_{3,39774}$=29.97, p<.001], and of Method [$F_{3,39746}$=13.49, p<.001] and of Method*Period [$F_{5,39777}$=4.00, p<.001] on Uncorrected Error Rate. Participants had higher error rates during PasS and ExpC (Table 1). Due to the intent estimation algorithm, error rates for these methods are artificially inflated in free typing as any word that is not in the dictionary is considered an error (e.g., abbreviations, words in other languages). Still, pairwise comparisons revealed significant differences on Total and Corrected Error Rates between all methods with the exception of PasS-ExpC. For Uncorrected Error Rates, the significant differences are between ExpC and all other methods (p<0.01), and ObsT and PasS *(p<0.05)*. There appear to be different correction behaviours depending on the method; furthermore, we found no significant differences in the Uncorrected Error Rate between experience sampling and passive sensing.

Multiple participants described how they were concerned about making errors during the prompted tasks and how they made an additional effort when compared to their normal typing behaviours.

5.2.3 Touch Dynamics. We found a significant main effect of Method [$F_{3,39763}$=55.47, p<.001] of Method*Period on Flight Time [$F_{5,39763}$=2.39, p<.05] and Hold time [$F_{3,39766}$=7.58, p<.001]. Pairwise comparisons revealed significant differences between all methods in Flight Time (p<.01), except for ExpC and PasS. For hold time only between ObsT and both ExpC and PasS (p<.01). All measures point towards participants performing faster in ObsT, followed by ExpT, PasS, and lastly ExpC. Only ExpC shows a significant interaction of Method*Period (p<0.05) with users taking 16 to 124 longer during the morning, which may have been a result of the questions asked during that period.

5.2.4 Text-entry Behaviours. We found a significant effect of Method [$F_{3,39768}$=31.15, p<.001] but not of Period [$F_{2,39766}$=3.46, p=.055] on use of Suggestions. We also found a main effect of Method on the use of autocorrect [$F_{3,39772}$=38.59, p<.001]. Interestingly, pairwise comparisons revealed significant differences between Passive Sensing and all other methods (p<.01), with suggestions and autocorrect being used more often during passive sensing. The only additional significant difference found was between ExpC and ExpT for the use of suggestions (p<.001). Although participants had the opportunity to use suggestions and auto-correct, they relied less on those features in experience sampling tasks and observed transcriptions.

5.3 Perceived Privacy, Trust, and Acceptance

In this section, we characterise participants perceptions around privacy and acceptance. We compare the weekly questionnaire responses across methods, the debriefing questionnaire, and present the results from the qualitative analyses.

Method	WPM	TER(%)	UER	CER	FT(ms)	HT(ms)	Sugg	AutoC
PasS	41.4 (14.9)	10.3 (3.3)	2.3 (1.6)	8.0 (3.1)	417.3 (343.4)	86.9 (21.3)	0.88 (1.03)	0.81 (0.65)
ExpT	48.5 (17.7)	7.0 (3.7)	2.4 (2)	4.7 (2.2)	358.2 (339.6)	84.0 (21.0)	0.40 (0.55)	0.33 (0.33)
ExpC	36.3 (12.2)	12.5 (8.2)	4.3 (4.9)	8.2 (5.1)	438.3 (373.9)	87.8 (23.6)	0.72 (0.56)	0.51 (0.48)
ObsT	53.4 (19.3)	4.5 (2.8)	1.1 (1.1)	3.5 (2.4)	306.4 (243.9)	81.4 (21.2)	0.24 (0.41)	0.19 (0.24)

Method	Privacy	Effort	Acceptance Q1
PasS	3.88 (0.19)	1.42 (0.19)	85.2 (19.1)
ExpT	4.34 (0.23)	1.69 (0.21)	70.2 (22.6)
ExpC	3.96 (0.21)	2.08 (0.26)	84.1 (18.4)

Table 1: Typing performance and behaviours measures, mean values, with standard deviations in brackets: Words per Minute (WPM), Total Error Rate (TER, %), Uncorrected Error Rate (UER), Corrected Error Rate (CER), Flight Time (FT, ms), Hold Time (HT, ms), Suggestions (Sugg), Autocorrect (AutoC), Privacy, Effort, acceptance Q1 (i.e. How willing would you be to continue using this method?).

Table 2: Summary of the findings for everyday text-entry collection.

Compliance & Coverage

- With low effort tasks, participants complied with 72% and 89% of the experience sampling tasks.
- Scheduling choices can affect effort & obtrusiveness.
- Participants would not be willing to use experience sampling for long periods.
- PasS collected about 7 times more data per week than experience sampling methods.

Typing Performance & Behaviours

- Words per minute were affected by the collection method.
- Error rates (Total, Corrected and Uncorrected) were affected by the data collection method with interactions with Period.
- Flight and Hold Time were affected by Method and Period.
- Suggestions and autocorrect were used significantly more with passive sensing.
- Participants showed better performance on ObsT, followed by ExpT, PasS and last ExpC.

Perceived Privacy, Trust, and Acceptance

- Passive collection was seamless, and required no additional effort.
- The privacy-aware design was essential for willingness to participate.
- ExpC can be perceived as the most demanding and intrusive tasks.
- For experience sampling, effort is participant dependent.
- Trust & willingness to share depends on whom, purpose and transparency.

5.3.1 Effort & Obtrusiveness. A Friedman Test on the debriefing question about required effort showed a statistically significant difference depending on the method ($\chi^2(2) = 9.800$, $p < .0.01$, $W = .188$) with ExpT M=1.69 (SD=1.05), ExpC M=2.08 (SD=1.32) and PasS M=1.42 (SD=0.95) (less is better), and no significant postdoc comparisons. The weekly questionnaires regarding acceptability revealed a significant difference in one measure ("How willing would you be to continue to use this method") [$F_{2,53}=5.40$, $p=.007$], with pairwise comparisons showing a significant difference between ExpT and PasS($p=.005$), Table 1).

Passive collection was seamless. Participants were not required to make any additional effort and quickly forgot their typing behaviours were being measured during everyday typing. For the most part, the keyboard was indistinguishable from their default keyboards and thought of as something that could be used indefinitely. *"In the first week I didn't even realised it (about participating in the study)... I used the keyboard normally as I would any other"* - P14. However, whenever a sensitive conversation was taking place it served as a trigger to remember they were actually participating in the study. Participants did not report acting differently, but the realisation of being observed might have affected behaviours during these trials. *"For the most part I didn't even remember. Only when I was having more private conversations; that was the only stimuli to remember"* - P3.

Cognitive effort only in compositions. Participants reported how both experience sampling tasks were low effort. However, most participants considered ExpC to be the most demanding due to the cognitive efforts involved in answering the questions. *"I had to work harder to answer questions, obviously. I had to think... Not a lot, the questions were clear, and there wasn't much to answer. (...) I had a couple of writing blocks (...) While in transcriptions, it was: see, do. It was easier for me"* - P8.

Effort is participant dependent for experience sampling. While the majority considered transcriptions to be easier due to the low cognitive demand, two participants reported the opposite. In particular, one with self-identified dyslexia, and the other reported having to concentrate more to ensure they were transcribing correctly. *"The one which requires more effort is the transcription, because we have to make sure to do the same as what is on the screen"* - P5.

Scheduling & Notifications affect perceived effort & obtrusiveness. A common sentiment shared by the participants was how the tasks were not demanding. Moreover, how the convenience of the flexible schedule facilitated compliance. For one user with a particularly tight morning schedule, the demands to tasks completion came solely from scheduling. *"The effort was not a lot, it was rather the worry with transcribing and answering within the time frame"* - P2.

Participants reported that a more cumbersome aspect often came from the notifications and the realisation of yet another task to do,

rather than the task in itself. *"Ha... again another task, I didn't want to, but it was quick. It was more the laziness of the realization [the awareness that you have a commitment]..."* - P21.

For compositions, the flexible schedule created situations where questions were not ideal (e.g., asking a user what they ate for breakfast at 9 am while the user had not eaten), creating additional cognitive load.

5.3.2 Privacy. We did not find significant differences in any of the questions from the weekly privacy questionnaire between the collection methods. However, in the debriefing questionnaire when we asked participants to rate the privacy of each method from 1 to 5 (from Nothing to Total), a Friedman Test showed a statistically significant difference, $\chi^2(2) = 7.74$, $p < .05$, with a effect size of $W = .15$, and no significant pairwise comparisons, with ExpT M=4.35 (SD=1.16), ExpC M=3.96 (SD=1.11) and PasSM=3.88 (SD=0.99) . While transcriptions were perceived as the most private, compositions and passive sensing had similar privacy ratings.

No raw text collection was essential for passive sensing. Overall, participants felt comfortable using a keyboard that assessed their typing behaviours as the high privacy scores indicate. However, it is worth mentioning that participants were aware of Wildkey's "no raw content collection" policy. Ensuring we would not store or collect written text seems to have positively affected the privacy ratings of passive sensing. All participants mentioned how they were only comfortable if no written content was shared/stored. *"It all depends how the data collection is made, (...), and how the data is stored and where. If you collected the sentences I write, I might have not been comfortable."* - P4.

Compositions can be the most invasive. Some participants felt the compositions tasks were, at times, intrusive. Although questions were related to mundane activities or general tastes, questions related to physical activity or dietary habits were considered by some to be uncomfortable to answer. *"I felt my privacy was being much more compromised when there were open questions."* - P4.

Transcriptions are perceived as the most private. Transcriptions are the only method where participants do not create content; rather, they are given sentences to copy. Participants were not concerned about their privacy being compromised in transcription tasks. *"There is no doubt that transcriptions tasks are the one that guarantee more privacy. Then, the answers to the questions [composition]. The one that is the least private is free use. On the other hand, evidently, I was not concerned with complying with the tasks in the weeks where there was free use. I used my phone, and that was it."* - P2.

Willingness to share depends on the context. When presented with the scenario of using typing behaviours to monitor/diagnose medical conditions, participants were more willing to do prompted tasks for more extended periods and more often. Moreover, participants wanted to share as much information as possible with their clinicians, except for not wanting to share any raw textual content. *"Is there anything you wouldn't want to share with your clinicians regarding this data? On the contrary, the more, the better"* - P10.

5.3.3 Trust &Transparency. Privacy is intrinsically linked with trust and transparency. In this section, we discuss how users perceived the efforts made to ensure they were in control and aware of the data they were generating for the study.

Transparency led to users' confidence. Participants were given a template study, which detailed the data that would be collected during the study; during the study, they received weekly emails with the number of characters contributed to the research; lastly, participants received a debriefing report with an analysis of their typing behaviours. These reports ensured users knew the study protocol and what data was being extracted. Making the study protocol clear to participants was crucial for passive sensing, particularly for people to agree to participate in the study *"They are useful, it's important to receive feedback of what we are doing and of what you are collecting, what sort of information you have."* - P17.

Private mode conveys trust. Participants reported not using the private mode, but still considered it to be a vital feature in the keyboard M=85.86 (SD=21.38) (i.e., from 0 to 100, how much do you agree with "the private mode is important"). With one exception, all participants advocated keeping the feature even going as far as stating their willingness to use and trust the keyboard depended on it. *"With the nuance of having the private mode, to me, because ... it allows that for some conversations to be more private and not observed."* - P1 For one user, the availability of a private mode was a sign to distrust the keyboard. If privacy was ensured with the no raw collection policy, why would it be necessary to provide such a feature, and suggested its removal? *"The lock [private mode] creates a lot of distrust. We are confronted with the scandals from Facebook, WhatsApp (...), they state they don't collect, don't sell, and don't share, but we are fully aware that it happens"* - P6. P6 statements clearly demonstrate the concerns of participants about the current surveillance capitalism practices of big tech companies.

Who and where affects trust. The keyboard application was made available to all participants through the Android PlayStore. Prior to the briefing , we asked participants to install the app (otherwise, they would be guided during the first session). It is important to highlight that where the app is made available and who is given and gives access can signal the trustworthiness of the monitoring technology. *"The perception of rigour increases, depending with whom you are installing. If it is a clinician, a therapist (. . .), (now if), you install it at home from the PlayStore [available to everyone]. I mean Instagram is there, games, everything is there"* - P6.

5.3.4 Motivating Participants. Regardless of the collection method, people have to be willing to participate in the study. For experience sampling methods, they need to continue to engage with the tasks at the allocated periods. In this section, we discuss some of the considerations that arose throughout the study on motivating participants.

Contribution feedback. In addition to promoting trust, frequent updates (e.g., weekly reports) can contribute to users' motivation. *"It's an incentive, it's good. I mean, I think they are essential, these reports. The person feels accompanied and feels that it wasn't making the effort for nothing. (...) It's like a small treat, a cute little thing we receive, and it makes us happy, happy to be contributing to something. It works as an acknowledgement"* - P8.

We only provided performance feedback at the end of the study. Multiple participants highlighted how they wanted more from their weekly feedback to assess their behaviours in-between weeks. *"If I wrote faster or slower. If I wrote faster in the morning or slower; those type of behaviours"* - P10.

Comparison among users. Participants believed typing behaviours could provide similar feedback to what they are used to with their own activity trackers (e.g., Fitbit). They wished to use its metrics to compare with others and track their daily/weekly typing activities and maybe even set goals. During the study, we had multiple groups of users who knew each other and compared their number of contributed characters throughout the weeks. Others asked how they compared against other users. These suggestions go in line with previous work that provided people with personalised feedback in exchange for participation in behavioural studies [48].

Purpose can be key for engagement. During the potential clinical scenario presented during the interviews, multiple participants associated their willingness to use Wildkey with its potential for clinical monitoring. Furthermore, when asked about willingness for continual use, many associated it with having a disease or being older (i.e. less healthy) and requiring more close monitoring, signalling health benefits and purpose can affect participant willingness to comply with in the wild text-entry collection methods. *"Well, no.... with my age, I don't think it would do any difference ... I mean, if the keyboard was already in my device I could use it without any issue (...) But with time, it would be OK (...) If I was older and the data was helpful for my doctor I would use it then."* - P7.

6 DISCUSSION

In this section, we break down the tradeoffs between experience sampling and passive collection, answering our overarching research question. We focus the discussion on 1) Compliance and Coverage, 2) Typing Performance and Behaviours, and lastly, two topics on user experience regarding 3) Privacy, Trust, and Transparency, as well as 4) Effort and Motivation. In all sections, we discuss and compare the data collection methods ExpT, ExpC, and PasS.

6.1 Compliance and Coverage

Experience sampling allows us to control when the users receive a stimulus to write. However, it comes at the cost of having to create flexible schedules to promote compliance and coverage. In this study, despite giving a minimum 4-hour window to comply, participants did not complete all tasks. There were significant differences between the two experience sample methods with ExpT at 89% and ExpC at 71% compliance (Figure 4). Although compliance is relatively high, we must consider that our study protocol was designed not to be demanding. Participants were prompted three times a day, and took less than five minutes to complete all tasks. Future research should take into account that more demanding studies will likely have lower compliance rates. Despite Experience Sampling affording more control on the anticipated collection coverage, the additional effort imposed on participant schedules can negatively impact their compliance. When directly asked about willingness to use experience sampling methods, all participants mentioned prompts would have to be sporadic. Moreover, participants were not willing to use them for extended periods (e.g., a month). In addition, our qualitative findings suggest that 'days off' between collection periods may positively impact overall compliance *("Respecting the weekend is good, we are freer [to rest]"* - P8), suggesting to carefully consider participants work-life commitments and seeking ways to integrate them into in the wild study designs.

While PasS does not guarantee any data collection points, the coverage and amount of text-entry data provided by passive data collection are unmatched for any user who regularly uses their smartphone for daily communication. For users who write less often (e.g., P1 had an average of 4 trials a day), combining passive sensing with experience sampling could provide optimal solution. PasS alone raises concerns with respect to privacy, trust, and can result in highly unbalanced datasets between participants. Our qualitative results suggest that the recruitment and use of Wildkey was facilitated by its crucial feature of collecting no raw textual content. Further studies seeking to access raw textual content must strive to run analysis locally and ensure no textual data leaves (or is stored) in the device. Moreover, data collection policies should be clearly communicated and one should strive to provide access and control to users (e.g., accessible reports, pause/stop features). When relying on PasS, we make the tradeoff between having the potential for gathering large amounts of data an getting highly unbalanced data (e.g., participants with 340k chars vs. 4k). We also lose information about the intent and context of the typing sessions. Future research should consider novel interactive features that allow users to tag passive sensing data, allowing for a better understanding of users' experience. Overall, passive sensing has the potential for high frequency sampling, coverage, and compliance, which for certain research applications such as context-aware keyboards, biometrics, or clinical monitoring are the most adequate path forward.

6.2 Typing Performance & Behaviours

In Remote Observed Transcriptions, participants typed significantly faster and more accurately (i.e., less Total and Corrected Error Rates) than in all other methods. Interestingly, users refrained from using autocorrect and suggestions during the tasks. A possible explanation is that users wanted to control what was written while being *observed*. Thus, studies that seek to investigate how people use predictive features will benefit from making evaluations in the wild. Flight and Hold Time were also significantly affected by the data collection method; participants in ObsT took the least amount of time (corroborating the performance metrics). Studies that seek to rely on typing dynamics to monitor or assess behaviours should also strive to conduct evaluations in the wild as controlled evaluations are limited in their ecological validity.

The following method where participants showed the best performance in terms of speed and errors was ExpT. Similar to ObsT, we believe the immediacy of receiving a clear prompt to do a mechanical task that requires no authoring effort enabled users to perform better. As such, ExpT is suitable when the evaluations focus on motor assessment and want to minimise the cognitive effort.

ExpC had the lowest performance of all methods, with error rates resembling PasS collection. The indecisiveness of what to write during a trial, and the potential mismatch between the question and

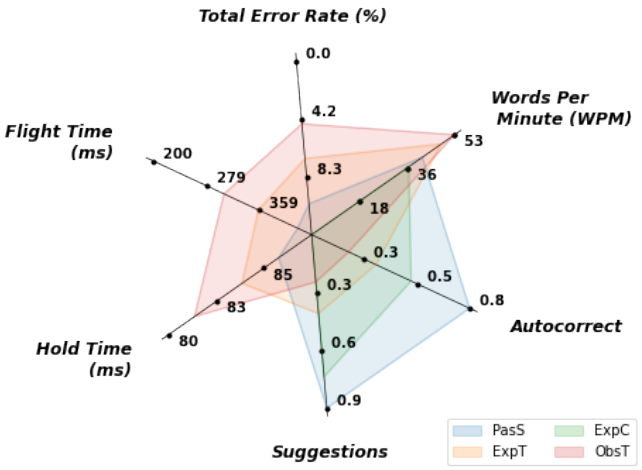

Figure 3: Tradeoff comparison plot. Performance, typing dynamics and behaviours comparison. Axes adapted for trends to represent "better" performance, with max values at the center for flight and hold time and total error rate.

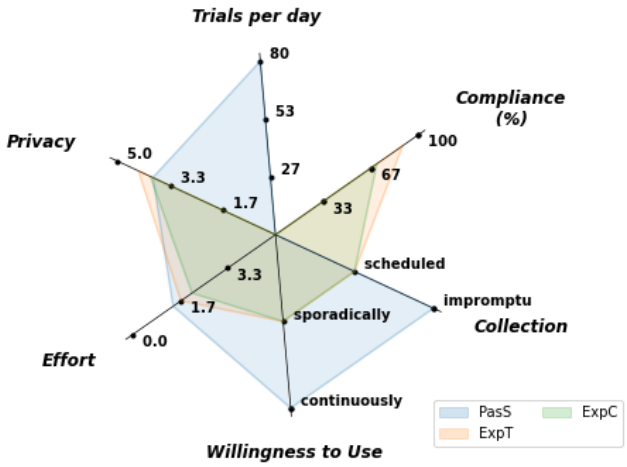

Figure 4: Tradeoff comparison plot. Compliance axis does not include PasS. Trials per Day Axis do not represent ExpC and ExpT since the protocol defines the trial limit. Effort axis is inverted as to represent higher trend less effort.

the user context may be artificially decreasing user performance. When we compare with PasS, it was as error-prone but slower, indicating that sampling through ExpC tasks might only be beneficial when we are interested in the cognitive load introduced, or want to increase it, or are interested in the responses typed, or want to be purposefully intrusive (e.g., questions about activity to promote healthy behavior).

Participants were slower in PasS than in ExpT and ObsT, contrasting with prior work [23, 43]. One reason might be the data selection process, which ensured only trials with 10 or more characters were analysed. Trials with fewer characters artificially increase speed due to the small sample size and require less cognitive effort to assemble.

Typing performance and behaviours are significantly different in the wild, and there are differences between (and within) experience sampling and passive sensing (Figure 3). Thus, one must carefully select the approach that best suits the context. Passive sensing represents more natural typing behaviours, but it introduces uncertainty in the data. Regarding performance, users are likely to be focused on the task during experience sampling, while there are no guarantees in passive sensing. Compositions introduce a cognitive effort and the need to author content, which resembles passive sensing in some metrics. Overall, composition tasks seem to be a compromise between transcriptions and passive sensing to collect everyday typing.

Lastly, Method and Period show significant interactions across multiple metrics (i.e., Flight Time, Total and Uncorrected Error rates), indicating that researchers should consider when to prompt users. Furthermore, this significant effect may indicate how behaviours and error rate metrics may be sensitive to other factors such as fatigue, which is in line with preliminary prior work [3] and worth further exploration.

6.3 Privacy, Trust & Transparency

We found a significant main effect of method on the privacy ratings in the debriefing questionnaire. While ExpT was considered the most private, ExpC shared similar ratings to PasS. Privacy perceptions in ExpC are highly dependent on the prompt, and personal, albeit mundane questions (i.e. "what did you have for lunch") were considered as intrusive as passive sensing. When using ExpC, one must carefully consider the questions and sampling schedule as both might affect the perceptions of privacy and intrusiveness of the method, which may lead to data quality problems (e.g., participants writing small sentences).

Overall, all methods had a positive privacy score; however, it is worth highlighting the steps taken by Wildkey and the protocol that may have had a positive effect. During the recruitment process, participants received a template report with example data of what would be collected. Furthermore, it was reinforced throughout recruitment communications that raw text content would not be collected under any circumstances. Based on our qualitative work, both these steps were crucial to ensure successful recruitment. During interviews, participants consistently mentioned how not collecting raw textual content is critical for their willingness to contribute. Interestingly, although participants would be happy to share more data for clinical purposes, they would still not be willing if it would require raw text collection. When collecting everyday typing data, one must ensure users are aware and perceive they control their data. In this study, we tackled this challenge through template reports, weekly feedback, final data report, and privacy control options (i.e., private mode), all of which received positive feedback, ranging from a must-have to nice to have from different participants. Future work could seek to provide further awareness and control by exploring review mechanisms that could potentially change what people are willing or not to contribute.

We used the final data report and interviews to encourage participants to reflect on their data sharing policies regarding typing

data and behaviours. It became evident during the interviews that purpose, context, and reward affect willingness to participate and share. Similarly, past work by Watson et al. [58] that explored smartphone data collection (both with consent/without) for public health emergencies highlighted considerations around location, consent mitigation, data reliability and privacy risks. When designing everyday text-entry collection methods, one must carefully consider what people are getting out of it and ensure they know how meaningful their contributions can be. We found success in relaying reports back to participants and presenting how typing data could be used as a monitoring tool for clinical purposes. For different contexts, the approach might vary from providing personal benefits (e.g., personalisation) to contribute to a greater good (e.g., data for language research).

Lastly, the identities to which the study/application are associated and the way they are made available appear to affect user trust in some capacity. Depending on the context, it might be essential to match onboarding and availability practices with users' expectations to maximise trust (e.g, a clinical application only be accessible through a 'prescription' special code).

6.4 Effort & Motivation

Experienced sampling, by its very nature, requires more effort from participants than passive collection. We found significant differences between the methods regarding perceived effort. ExpC required more effort due to the cognitive load of answering prompted questions, while ExpT was widely regarded as effortless. However, for a few participants, the demand for careful reading made it more stressful than answering open questions. Depending on the experience sampling schedule (i.e., frequency and duration), we can expect the differences to passive sensing to be exacerbated.

A downside to experience sampling is the need to rely on notifications to promote compliance. Unfortunately, notifications and scheduling will impact participants' perceived effort and cumbersomeness. While passive collection unassumingly hides in the background, notifications disrupt and create additional load for participants, adding another factor to the analysis and study design.

In experience sampling, one has to create the tasks to request. However, the content of the tasks may affect participants' motivation and perceived effort. Some participants highlighted how transcriptions were an enjoyable part of the day, learning about new proverbs, while others enjoyed answering some of the one-time questions about their hobbies. Our findings suggest that one could promote engagement and reduce perceived effort if users could personalise the type of content of experience sampling prompts.

Feedback on participants' performance and contribution can also promote engagement. In this study, participants received a weekly report (i.e., "You contributed with X characters. Thank You!") that served different purposes for different users. For some, it was used to compare their contribution with other friends/colleagues who were also partaking in the study. For others, it was a weekly reminder that everything was running smoothly. Lastly, others took it as a small appreciation for their effort. While a simple weekly report was enough for most, some requested additional information such as performance metrics or benchmarks to the average keyboard user. Our results call attention to the benefits of providing regular feedback to users to promote engagement, possibly affecting perceived effort and compliance.

Despite the unassuming differences in the effort score, participants overwhelmingly shared that they would not be willing to use experience sampling methods in the long term. Experience sampling methods seem reasonable for limited time frames and with a clear purpose (e.g., conducting a clinical evaluation). However, for PasS, if the transition to a new keyboard is seamless, and privacy concerns are addressed (i.e., no raw content and control over sharing), participants believed they could use it indefinitely.

6.5 Limitations and Future Work

For most participants, Wildkey was indistinguishable from their keyboard. However, the keyboard lacks some features that make the transition more noticeable to some. Currently, it does not support swipe typing, multi-language support, nor does it provide emojis, all of which can be essential to further our understanding of text-entry behaviours and highly relevant topics for future work. For some participants, this meant they had to change their behaviours to adapt to the keyboard features (e.g., using app emojis instead of keyboard emojis). Our results suggest that not collecting raw text during passive sensing was crucial for recruitment, compliance, and capturing natural typing behaviours. However, this forced all analysis to be performed on the device. Our approach does not allow for text reconstruction, preventing new metrics to be computed after the data collection stage. Wildkey could not deal with multiple cursor changes during passive sensing, which should be investigated in future work. The use of an intent prediction algorithm for passive sensing and compositions can impact error rates, artificially increasing them. For instance, all abbreviations or emphasis in words (e.g., "nooo") will be considered errors as they do not exist in the dictionary. We purposefully did not collect when/how the private mode was used, nor in which applications participants were typing. Although the approach is more respectful of users privacy, it inevitably restricts the analysis one can due. The WildKey toolkit will be further developed to ensure it can support less strict privacy concerns, without jeopardising user privacy and confidence. Lastly, we focus this work on trials with at least 10 characters, as shorter trials would not allow for accurate comparisons with other experience sampling methods. Nevertheless, short trials are common, and potentially interesting.

7 CONCLUSION

Everyday text-entry collection from real world settings is a contested space. While these datasets hold untold opportunity for researchers in mobile interaction design, linguists, biometrics and digital health, they are simultaneously viewed as highly sensitive and private information by the device users. Prior studies have commonly adopted one of two distinct strategies when approaching everyday text-entry collection, Experience Sampling with frequent prompts and requests for participants to complete prescribed text input tasks; and Passive Sensing of natural device usage through invasive or restrictive logging of all user interactions with the keyboard. Both methods require a degree of compromise with respect to study design or participant acceptance and compliance, yet no prior

works have reported on the tradeoffs of the text-entry collection methods.

In this paper we present Wildkey, an experimental research platform, including a mobile keyboard, to support in the wild text-entry data collection using a combination of Experience Sampling and Passive Sensing. Through a four week study with 26 participants we investigated the tradeoffs of text-entry collections methods. We reported on our empirical evaluation that quantitatively and qualitatively compares the compliance & data collection coverage, user's typing performance, and user experience & acceptability of these two methods. Our results revealed that everyday text-entry collection methods not only had a significant effect on the typing dynamics and behaviours, but also impacted on the participants willingness to engage or trust research studies adopting these methods.

Through this work we offer novel understandings of the tradeoffs between data collection methods that can inform the future design of everyday text-entry studies, and provide new guidance and practical approaches to improve the user acceptability of collecting data in-situ.

ACKNOWLEDGMENTS

This project was partially supported by FCT through LASIGE Research Unit funding, ref. UIDB/00408/2020, LARSyS Research Unit funding, ref. UIDB/50009/2020, project mIDR (AAC 02/SAICT/-2017, project 30347, cofunded by COMPETE/FEDER/FNR), and the IDEA-FAST project which has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No. 853981. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA and associated partners.

REFERENCES

[1] 2021. Android Open Source Project. https://source.android.com/. (Accessed on 05/18/2021).

[2] Hilal Al-Libawy, Ali Al-Ataby, Waleed Al-Nuaimy, and Majid A. Al-Taee. 2017. Enhanced operator fatigue detection method based on computer-keyboard typing style. In *2017 14th International Multi-Conference on Systems, Signals Devices (SSD)*. 217–221. https://doi.org/10.1109/SSD.2017.8166991

[3] Hilal Al-Libawy, Ali Al-Ataby, Waleed Al-Nuaimy, Majid A. Al-Taee, and Qussay Al-Jubouri. 2016. Fatigue Detection Method Based on Smartphone Text Entry Performance Metrics. In *2016 9th International Conference on Developments in eSystems Engineering (DeSE)*. 40–44. https://doi.org/10.1109/DeSE.2016.9

[4] Kenneth C. Arnold, Krzysztof Z. Gajos, and Adam T. Kalai. 2016. On Suggesting Phrases vs. Predicting Words for Mobile Text Composition. *Proceedings of the 29th Annual Symposium on User Interface Software and Technology - UIST '16* (2016), 603–608. https://doi.org/10.1145/2984511.2984584

[5] Teresa Arroyo-Gallego, María Jesus Ledesma-Carbayo, Alvaro Sánchez-Ferro, Ian Butterworth, Carlos S Mendoza, Michele Matarazzo, Paloma Montero, Roberto López-Blanco, Veronica Puertas-Martin, Rocio Trincado, and Luca Giancardo. 2017. Detection of Motor Impairment in Parkinson's Disease Via Mobile Touchscreen Typing. *IEEE Transactions on Biomedical Engineering* 64, 9 (2017), 1994–2002. https://doi.org/10.1109/TBME.2017.2664802

[6] L Barnard, J S Yi, J A Jacko, and A Sears. 2007. Capturing the effects of context on human performance in mobile computing systems. *Personal and Ubiquitous Computing* 11, 2 (2007), 81–96.

[7] Michael Beißwenger and Angelika Storrer. 2008. Corpora of computer-mediated communication. *Corpus Linguistics. An International Handbook. Series: Handbücher zur Sprach-und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science. Mouton de Gruyter, Berlin* (2008).

[8] Florian Bemmann and Daniel Buschek. 2020. LanguageLogger: A Mobile Keyboard Application for Studying Language Use in Everyday Text Communication in the Wild. *Proc. ACM Hum.-Comput. Interact.* 4, EICS, Article 84 (June 2020), 24 pages. https://doi.org/10.1145/3397872

[9] Xiaojun Bi, Tom Ouyang, and Shumin Zhai. 2014. Both Complete and Correct? Multi-Objective Optimization of Touchscreen Keyboard. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2297–2306. https://doi.org/10.1145/2556288.2557414

[10] Xiaojun Bi and Shumin Zhai. 2016. *IJQwerty: What Difference Does One Key Change Make? Gesture Typing Keyboard Optimization Bounded by One Key Position Change from Qwerty*. Association for Computing Machinery, New York, NY, USA, 49–58. https://doi.org/10.1145/2858036.2858421

[11] Barry Brown, Stuart Reeves, and Scott Sherwood. 2011. *Into the Wild: Challenges and Opportunities for Field Trial Methods*. Association for Computing Machinery, New York, NY, USA, 1657–1666. https://doi.org/10.1145/1978942.1979185

[12] Ulrich Burgbacher and Klaus Hinrichs. 2014. An Implicit Author Verification System for Text Messages Based on Gesture Typing Biometrics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '14)*. Association for Computing Machinery, New York, NY, USA, 2951–2954. https://doi.org/10.1145/2556288.2557346

[13] Daniel Buschek, Benjamin Bisinger, and Florian Alt. 2018. *ResearchIME: A Mobile Keyboard Application for Studying Free Typing Behaviour in the Wild*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173829

[14] Daniel Buschek, Alexander De Luca, and Florian Alt. 2016. *Evaluating the Influence of Targets and Hand Postures on Touch-Based Behavioural Biometrics*. Association for Computing Machinery, New York, NY, USA, 1349–1361. https://doi.org/10.1145/2858036.2858165

[15] Scott Carter, Jennifer Mankoff, Scott R Klemmer, and Tara Matthews. 2008. Exiting the Cleanroom: On Ecological Validity and Ubiquitous Computing. *Human–Computer Interaction* 23, 1 (2008), 47–99. https://doi.org/10.1080/07370020701851086

[16] Matteo Ciman, Katarzyna Wac, and Ombretta Gaggi. 2015. iSensestress: Assessing stress through human-smartphone interaction analysis. In *2015 9th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. 84–91. https://doi.org/10.4108/icst.pervasivehealth.2015.259280

[17] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. *Observations on Typing from 136 Million Keystrokes*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174220

[18] Neil Dhir, Mathias Edman, Álvaro Sanchez Ferro, Tom Stafford, and Colin Bannard. 2020. Identifying robust markers of Parkinson's disease in typing behaviour using a CNN-LSTM network. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Online, 578–595. https://doi.org/10.18653/v1/2020.conll-1.47

[19] Benjamin Draffin, Jiang Zhu, and Joy Zhang. 2014. KeySens: Passive User Authentication through Micro-behavior Modeling of Soft Keyboard Interaction. In *Mobile Computing, Applications, and Services*, Gérard Memmi and Ulf Blanke (Eds.). Springer International Publishing, Cham, 184–201.

[20] Michelle Drouin and Brent Driver. 2014. Texting, textese and literacy abilities: a naturalistic study. *Journal of Research in Reading* 37, 3 (2014), 250–267. https://doi.org/10.1111/j.1467-9817.2012.01532.x

[21] Mark Dunlop and John Levine. 2012. *Multidimensional Pareto Optimization of Touchscreen Keyboards for Speed, Familiarity and Improved Spell Checking*. Association for Computing Machinery, New York, NY, USA, 2669–2678. https://doi.org/10.1145/2207676.2208659

[22] Clayton Epp, Michael Lippold, and Regan L. Mandryk. 2011. *Identifying Emotional States Using Keystroke Dynamics*. Association for Computing Machinery, New York, NY, USA, 715–724. https://doi.org/10.1145/1978942.1979046

[23] Abigail Evans and Jacob Wobbrock. 2012. Taming wild behavior: the input observer for obtaining text entry and mouse pointing measures from everyday computer use. In *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 1947–1956.

[24] Leah Findlater and Jacob Wobbrock. 2012. *Personalized Input: Improving Ten-Finger Touchscreen Typing through Automatic Adaptation*. Association for Computing Machinery, New York, NY, USA, 815–824. https://doi.org/10.1145/2207676.2208520

[25] Mayank Goel, Leah Findlater, and Jacob Wobbrock. 2012. *WalkType: Using Accelerometer Data to Accomodate Situational Impairments in Mobile Touch Screen Text Entry*. Association for Computing Machinery, New York, NY, USA, 2687–2696. https://doi.org/10.1145/2207676.2208662

[26] Jennifer Golbeck, Cristina Robles, Michon Edmondson, and Karen Turner. 2011. Predicting Personality from Twitter. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. 149–156. https://doi.org/10.1109/PASSAT/SocialCom.2011.33

[27] J Goodman, G Venolia, K Steury, and C Parker. 2002. Language modeling for soft keyboards. In *Proceedings of the 7th international conference on Intelligent user interfaces*. ACM, 194–195.

[28] João Guerreiro, André Rodrigues, Kyle Montague, Tiago Guerreiro, Hugo Nicolau, and Daniel Gonçalves. 2015. TabLETS Get Physical: Non-Visual Text Entry on Tablet Devices. In *Proceedings of the 33rd Annual ACM Conference on Human*

Factors in Computing Systems (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 39–42. https://doi.org/10.1145/2702123.2702373

[29] Asela Gunawardana, Tim Paek, and Christopher Meek. 2010. Usability Guided Key-Target Resizing for Soft Keyboards. In *Proceedings of the 15th International Conference on Intelligent User Interfaces* (Hong Kong, China) *(IUI '10)*. Association for Computing Machinery, New York, NY, USA, 111–118. https://doi.org/10.1145/1719970.1719986

[30] Jonathan Gurary, Ye Zhu, Nahed Alnahash, and Huirong Fu. 2016. Implicit Authentication for Mobile Devices Using Typing Behavior. In *Human Aspects of Information Security, Privacy, and Trust*, Theo Tryfonas (Ed.). Springer International Publishing, Cham, 25–36.

[31] Niels Henze, Enrico Rukzio, and Susanne Boll. 2012. Observational and experimental investigation of typing behaviour using virtual keyboards for mobile devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 2659–2668. https://doi.org/10.1145/2207676.2208658

[32] Susan C Herring and John C Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics* 10, 4 (2006), 439–459. https://doi.org/10.1111/j.1467-9841.2006.00287.x

[33] John Hunsley. 1992. Development of the Treatment Acceptability Questionnaire. *Journal of Psychopathology and Behavioral Assessment* 14, 1 (01 Mar 1992), 55–64. https://doi.org/10.1007/BF00960091

[34] Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, Sevasti Bostantjopoulou, Zoe Katsarou, Lisa Klingelhoefer, Heinz Reichmann, Sofia B Dias, José A Diniz, Dhaval Trivedi, K Ray Chaudhuri, and Leontios J Hadjileontiadis. 2018. Motor Impairment Estimates via Touchscreen Typing Dynamics Toward Parkinson's Disease Detection From Data Harvested In-the-Wild. *Frontiers in ICT* 5 (2018), 28. https://doi.org/10.3389/fict.2018.00028

[35] Hassan Khan, Aaron Atwater, and Urs Hengartner. 2014. A Comparative Evaluation of Implicit Authentication Schemes. In *Research in Attacks, Intrusions and Defenses*, Angelos Stavrou, Herbert Bos, and Georgios Portokalidis (Eds.). Springer International Publishing, Cham, 255–275.

[36] Andreas Komninos, Mark Dunlop, Kyriakos Katsaris, and John Garofalakis. 2018. A Glimpse of Mobile Text Entry Errors and Corrective Behaviour in the Wild. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, 221–228. https://doi.org/10.1145/3236112.3236143

[37] K H Lam, K A Meijer, F C Loonstra, E M E Coerver, J Twose, E Redeman, B Moraal, F Barkhof, V de Groot, B M J Uitdehaag, and J Killestein. 2021. Real-world keystroke dynamics are a potentially valid biomarker for clinical disability in multiple sclerosis. *Multiple Sclerosis Journal* 27, 9 (2021), 1421–1431. https://doi.org/10.1177/1352458520968797

[38] Chae Young Lee, Seong Jun Kang, Sang-Kyoon Hong, Hyeo-Il Ma, Unjoo Lee, and Yun Joong Kim. 2016. A Validation Study of a Smartphone-Based Finger Tapping Application for Quantitative Assessment of Bradykinesia in Parkinson's Disease. *PLOS ONE* 11, 7 (07 2016), 1–11. https://doi.org/10.1371/journal.pone.0158852

[39] I. Scott MacKenzie, Tatu Kauppinen, and Miika Silfverberg. 2001. Accuracy Measures for Evaluating Computer Pointing Devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 9–16. https://doi.org/10.1145/365024.365028

[40] I. Scott MacKenzie and R. William Soukoreff. 2002. Text Entry for Mobile Computing: Models and Methods,Theory and Practice. *Human–Computer Interaction* 17, 2-3 (2002), 147–198. https://doi.org/10.1080/07370024.2002.9667313 arXiv:https://www.tandfonline.com/doi/pdf/10.1080/07370024.2002.9667313

[41] Alex Mariakakis, Sayna Parsi, Shwetak N. Patel, and Jacob O. Wobbrock. 2018. Drunk User Interfaces: Determining Blood Alcohol Level Through Everyday Smartphone Tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 234, 13 pages. https://doi.org/10.1145/3173574.3173808

[42] Charles E. McCulloch and John M. Neuhaus. 2011. Prediction of Random Effects in Linear and Generalized Linear Models under Model Misspecification. *Biometrics* 67, 1 (2011), 270–279. https://doi.org/10.1111/j.1541-0420.2010.01435.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2010.01435.x

[43] Hugo Nicolau, Kyle Montague, Tiago Guerreiro, André Rodrigues, and Vicki L Hanson. 2017. Investigating Laboratory and Everyday Typing Performance of Blind Users. *ACM Transactions on Accessible Computing (TACCESS)* 10, 1 (mar 2017), 4:1—-4:26. https://doi.org/10.1145/3046785

[44] Antti Oulasvirta, Anna Reichel, Wenbin Li, Yan Zhang, Myroslav Bachynskyi, Keith Vertanen, and Per Ola Kristensson. 2013. *Improving Two-Thumb Text Entry on Touchscreen Devices*. Association for Computing Machinery, New York, NY, USA, 2765–2774. https://doi.org/10.1145/2470654.2481383

[45] Leysia Palen and Marilyn Salzman. 2002. Voice-Mail Diary Studies for Naturalistic Data Capture under Mobile Conditions. In *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work* (New Orleans, Louisiana, USA) *(CSCW '02)*. Association for Computing Machinery, New York, NY, USA, 87–95. https://doi.org/10.1145/587078.587092

[46] Kseniia Palin, Anna Maria Feit, Sunjun Kim, Per Ola Kristensson, and Antti Oulasvirta. 2019. How do people type on mobile devices? Observations from a study with 37,000 volunteers. *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services, MobileHCI 2019* (2019). https://doi.org/10.1145/3338286.3340120

[47] Chang Sup Park and Barbara K. Kaye. 2019. Smartphone and self-extension: Functionally, anthropomorphically, and ontologically extending self via the smartphone. *Mobile Media & Communication* 7, 2 (2019), 215–231. https://doi.org/10.1177/2050157918808327 arXiv:https://doi.org/10.1177/2050157918808327

[48] Katharina Reinecke and Krzysztof Z Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1364–1378. https://doi.org/10.1145/2675133.2675246

[49] Shyam Reyal, Shumin Zhai, and Per Ola Kristensson. 2015. *Performance and User Experience of Touchscreen and Gesture Keyboards in a Lab Setting and in the Wild*. Association for Computing Machinery, New York, NY, USA, 679–688. https://doi.org/10.1145/2702123.2702597

[50] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E P Seligman, and Lyle H Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE* 8, 9 (2013), 1–16. https://doi.org/10.1371/journal.pone.0073791

[51] Beat Siebenhaar. 2006. Code choice and code-switching in Swiss-German Internet Relay Chat rooms. *Journal of Sociolinguistics* 10, 4 (2006), 481–506. https://doi.org/10.1111/j.1467-9841.2006.00289.x

[52] R William Soukoreff and I Scott MacKenzie. 2003. Metrics for text entry research: an evaluation of MSD and KSPC, and a new unified error metric. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 113–120. https://doi.org/10.1145/642611.642632

[53] Clemens Stachl, Sven Hilbert, Jiew–Quay Au, Daniel Buschek, Alexander De Luca, Bernd Bischl, Heinrich Hussmann, and Markus Bühner. 2017. Personality Traits Predict Smartphone Usage. *European Journal of Personality* 31, 6 (2017), 701–722. https://doi.org/10.1002/per.2113

[54] Pin Shen Teh, Ning Zhang, Andrew Beng Jin Teoh, and Ke Chen. 2016. A survey on touch dynamics authentication in mobile devices. *Computers & Security* 59 (2016), 210–235. https://doi.org/10.1016/j.cose.2016.03.003

[55] Mindaugas Ulinskas, Robertas Damaševičius, Rytis Maskeliūnas, and Marcin Woźniak. 2018. Recognition of human daytime fatigue using keystroke data. *Procedia Computer Science* 130 (2018), 947–952. https://doi.org/10.1016/j.procs.2018.04.094

[56] Lieke Verheijen and Wessel Stoop. 2016. Collecting Facebook Posts and WhatsApp Chats. In *Text, Speech, and Dialogue*, Petr Sojka, Aleš Horák, Ivan Kopeček, and Karel Pala (Eds.). Springer International Publishing, Cham, 249–258.

[57] Keith Vertanen and Per Ola Kristensson. 2014. Complementing Text Entry Evaluations with a Composition Task. *ACM Trans. Comput.-Hum. Interact.* 21, 2 (feb 2014). https://doi.org/10.1145/2555691

[58] Colin Watson, Ridita Ali, and Jan David Smeddinck. 2021. Tensions and Mitigations: Understanding Concerns and Values around Smartphone Data Collection for Public Health Emergencies. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 330 (oct 2021), 31 pages. https://doi.org/10.1145/3476071

[59] Daryl Weir, Simon Rogers, Roderick Murray-Smith, and Markus Löchtefeld. 2012. A User-Specific Machine Learning Approach for Improving Touch Accuracy on Mobile Devices. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12)*. Association for Computing Machinery, New York, NY, USA, 465–476. https://doi.org/10.1145/2380116.2380175

[60] Jacob O. Wobbrock and Brad A. Myers. 2006. Analyzing the Input Stream for Character- Level Errors in Unconstrained Text Entry Evaluations. *ACM Trans. Comput.-Hum. Interact.* 13, 4 (Dec. 2006), 458–489. https://doi.org/10.1145/1188816.1188819

[61] Tal Yarkoni. 2010. Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality* 44, 3 (2010), 363–373. https://doi.org/10.1016/j.jrp.2010.04.001

[62] Ying Yin, Tom Yu Ouyang, Kurt Partridge, and Shumin Zhai. 2013. Making Touchscreen Keyboards Adaptive to Keys, Hand Postures, and Individuals: A Hierarchical Spatial Backoff Model Approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2775–2784. https://doi.org/10.1145/2470654.2481384

[63] S Zhai, M Hunter, and B A Smith. 2002. Performance optimization of virtual keyboards. *Human–Computer Interaction* 17, 2-3 (2002), 229–269.

[64] Mingrui Zhang and Jacob O. Wobbrock. 2019. Beyond the input stream: Making text entry evaluations more flexible with transcription sequences. *UIST 2019 - Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (2019), 831–842. https://doi.org/10.1145/3332165.3347922

[65] Mingrui "Ray" Zhang, Alex Mariakakis, Jacob Burke, and Jacob O. Wobbrock. 2021. A Comparative Study of Lexical and Semantic Emoji Suggestion Systems. In *Diversity, Divergence, Dialogue*, Katharina Toeppe, Hui Yan, and Samuel Kai Wah

Chu (Eds.). Springer International Publishing, Cham, 229–247.
[66] Mingrui Ray Zhang and Shumin Zhai. 2021. *PhraseFlow: Designs and Empirical Studies of Phrase-Level Input.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445166